Prof. Dr. Budiyono, M.Sc.

PENGANTAR PENILAIAN HASIL BELAJAR

SEBELAS MARET UNIVERSITY PRESS

Perpustakaan Nasional : Katalog Dalam Terbitan (KDT)

Prof. Dr. Budiyono, M.Sc.
Pengantar Penilaian Hasil Belajar. Cetakan 1 . Surakarta . UPT UNS Press. 2015
x + 202 hal; 24.5 cm

PENGANTAR PENILAIAN HASIL BELAJAR

Penulis
Prof. Dr. Budiyono, M.Sc.
Editor
Drs. Suyono, M.Si.

Ilustrasi Sampul
UPT UNS Press
Penerbit
UPT Penerbitan dan Pencetakan UNS (UNS Press)
Jl. Ir. Sutami No. 36 A Kentingan, Jawa Tengah, Indonesia 57126
Telp. 0271-646994 Psw. 341 Faximale 0271-7890628
Website : www.unspress.uns.ac.id
Email: unspress@uns.ac.id

Cetakan 1, Edisi 1, Januari 2015

ISBN 978-979-498-958-6

PRAKATA

Buku ini ditulis dengan tujuan untuk menambah pustaka buku-buku yang berbicara mengenai penilaian berbahasa Indonesia. Sasaran buku ini adalah para peneliti, khususnya, para peneliti pemula bidang pendidikan dan psikologi, dan para guru yang telah mengajar di lapangan untuk menambah bekal pengetahuan mengenai penilaian. Tentu saja buku ini dapat dipakai sebagai referensi bagi mahasiswa yang sedang menempuh matakuliah penilaian hasil belajar.

Buku ini terdiri dari sebelas bab. Bab I membahas konsep pengukuran dan penilaian. Bab II memuat secara ringkas mengenai teori pengukuran, baik teori tes klasik maupun teori respons butir. Pada bab ini dibicarakan asumsi-asumsi pada teori tes klasik dan teorema-teorema yang mengikutinya. Bagi pembaca yang tidak terlalu menyukai matematika, bab ini dapat dilewati. Bab III memuat tes dan persyaratannya, yaitu validitas dan reliabilitas. Bab IV mendiskusikan berbagai hal mengenai penilaian pada ranah kognitif. Pada bab ini dibicarakan mengenai berbagai jenis tes tertulis, baik untuk *constructed-response test* maupun *selected-response test*, termasuk tes pilihan ganda. Bab V membicarakan analisis butir pada penilaian ranah kognitif, termasuk pembicaraan mengenai tingkat kesuliran, daya pembeda, maupun berfungsinya pengecoh. Bab VI membiacarakan non tes, di antaranya adalah pengukuran sikap dengan skala Likert, skala Thurstone, dan skala beda semantik. Bab VII membicarakan instrumen penilaian untuk ranah afektif, cara penyusunannya, validitas, dan reliabilitasnya. Bab VIII mendiskusikan instrumen penilaian pada ranah psikomotor beserta cara menyusunnya. Bab IX memuat beberapa penilaian alternatif, yaitu penilaian berbasis kelas, penilaian untuk pembelajaran, dan penilaian otentik. Bab X memut penilaian portofolio, penilaian yang mulai mendapatkan perhatian di sekolah. Bab terakhir, yaitu Bab XI memuat hal yang belum

banyak diketahui orang, yaitu mengenai bias butir atau keberbedaan fungsi butir (*differential item functioning*).

Kepada semua pihak yang membantu dan memungkinkannya diterbitkannya buku ini, terutama kepada Sebelas Maret University Press, penulis mengucapkan terima kasih. Mudah-mudahan usaha kecil ini bermanfaat besar. Tak ada gading yang tak retak, saran dan kritik membangun dari pembaca akan penulis terima dengan senang hati.

Surakarta, Januari 2015
Budiyono

DAFTAR ISI

BAB I
PENGUKURAN DAN PENILAIAN

PENDAHULUAN

Pendidik yang baik seharusnya berkeinginan untuk mengetahui apakah hal-hal yang disampaikan di kelas dapat diterima dengan baik oleh peserta didiknya atau tidak. Pendidik yang baik seharusnya ingin tahu apakah peserta didiknya telah belajar pada arah yang benar atau tidak. Pendidik yang baik juga pasti berkeinginan untuk membantu peserta didik yang mengalami kesulitan belajar. Untuk mengetahui hal-hal seperti itulah diperlukan apa yang disebut asesmen (*assessment*) atau penilaian. [1]

Dulu orang beranggapan bahwa pembelajaran dan penilaian adalah dua kegiatan yang terpisah. Namun demikian, sekarang berkembang "pandangan baru" mengenai pembelajaran, bahwa "*when you teach, you begin with assessment*" (DiRanna, et al. 2008: 7). Ini berarti, pembelajaran dan penilaian adalah dua kegiatan yang saling berintegrasi, tak terpisahkan, dari awal sampai akhir pembelajaran.

PENGUKURAN

Untuk dapat melakukan penilaian, dilakukan suatu kegiatan yang disebut pengukuran (*measurement*).

Allen dan Yen (1979: 2) mendefinisikan pengukuran sebagai pemberian bilangan kepada seseorang dengan cara yang sistematis untuk menyatakan sifat-sifat seseorang (*measurement is the assigning of numbers to individuals in a systematic way as a means of representing properties of*

---

[1] Pada buku ini istilah asesmen dan penilaian dianggap istilah yang sama. Kadang disebut dengan istilah asesmen, kadang disebut dengan istilah penilaian.

the individuals). Di sisi lain. Reynolds. Livingstone. dan Willson (2010: 3) mendefinisikan "*measurement is a set of rules for assigning numbers to represent objects, traits, attributes, or behaviors*". Pada definisi tersebut. pengukuran diartikan sebagai sekumpulan cara untuk memberikan bilangan kepada objek. kemampuan. atribut. atau perilaku untuk menyatakan kuantitas objek, kemampuan. atribut atau perilaku yang diukur. Bilangan-bilangan yang dilekatkan sebagai hasil pengukuran harus dilakukan dengan proses yang diatur secara cermat. hati-hati, dan dapat diulang (*repeatable procedure*).

Teori pengukuran adalah cabang statistik terapan yang berusaha untuk: (1) menjelaskan. mengkategorisasi, dan mengevaluasi kualitas pengukuran. (2) meningkatkan kegunaan. akurasi. dan kebermaknaan pengukuran. dan (3) mengembangkan metodologi untuk pengembangan instrumen (*instrument development*) pengukuran yang baru dan lebih baik.

Sejarah teori pengukuran dapat dibaca dari beberapa buku, misalnya dari buku yang berjudul *A History of Psychological Testing* yang ditulis oleh Philip H. Du Bois. pada 1970 (Allen dan Yen, 1979: 2). Isi dari buku itu dapat disarikan sebagai berikut.

Teori pengukuran meningkatkan penggunaan ujian dalam tiga kelompok besar bidang pengembangan, yaitu: (1) ujian-ujian kepada pega-wai (*civil service exam*), (2) ujian-ujian sekolah (*school exams*), dan (3) studi mengenai perbedaan individual (*study of individual differences*). Ujian penerimaan pegawai dimulai di China kira-kira tiga ribu tahun yang lalu ketika kerajaan memerlukan pengetahuan untuk mengukur kompetensi pegawainya. Sampai pada abad kedua belas, siswa-siswa di sekolah-sekolah di Eropa diberi ujian lisan. Setelah dapat diciptakan kertas ujian-ujian tulis mulai diberlakukan.

Studi mengenai perbedaan individu dimulai di Inggris ketika Sir Francis Galton (1822–1911) mendirikan laboratorium terkenal yang disebut *Antrometric Laboratory* yang berisi instrumen untuk mengukur berbagai ketrampilan sensori dan gerak motorik (*sensory and motor skills*). Karl Pearson (1857–1936) mengembangkan berbagai teknik statistik sebagai inti dari dasar-dasar teori pengukuran. Di Perancis, Alfred Binet (1857–1911) mengembangkan pertama kali tes inteligensi pada tahun 1905 sebagai bagian dari studinya mengenai perbedaan individu. Di Jerman. William Stern (1871–1938) mengembangkan tingkat kecerdasan (*intelligence quotien*, IQ) yang didefinisikan sebagai perbandingan antara *mental age* (yang diukur) dengan *chronological age* (yang senyatanya). Di Inggris. Charles Spearman (1863–1945), sebagai pengikut Galton dan Pearson, mengembangkan berbagai cara untuk mengukur koefisien reliabilitas.

Mula-mula tes (ujian) diperuntukkan secara individual, *one individual at a time*. Ujian secara klasikal muncul pertama kali ketika Amerika Serikat memberikan ujian kepada calon-calon pasukan militer pada Perang Dunia Pertama. Kesuksesan Amerika Serikat menggunakan ujian klasikal menjadi pemicu digunakannya ujian klasikal di sekolah-sekolah dan industri, sampai dengan saat ini.

Buku-buku mengenai teori pengukuran sudah lama ditulis orang. Misalnya, pada tahun 1904, E. L. Thorndike menulis buku yang berjudul *An Introduction to the Theory of Mental and Social Measurement*. Walaupun sudah dibicarakan lama, namun teori pengukuran sebagai salah satu cabang ilmu mulai berkembang secara serius sekitar tahun 1930-an. Mulai pada masa itu, jurnal-jurnal ilmiah mengenai teori pengukuran mulai bermunculan. Pada tahun 1935, diterbitkan jurnal yang berbicara mengenai teori pengukuran yang diberi nama jurnal *Psychometrica*. Pada 1941 terbit jurnal *Educational and Psychological Measurement*. Pada tahun 1947 terbit jurnal *British Journal of Statistical Psychology*. Jurnal-jurnal mengenai teori pengukuran sekarang ini sudah sangat banyak. Penelitian-penelitian mengenai teori pengukuran terus berkembang sampai sekarang

Untuk melakukan pengukuran perlu menggunakan alat ukur. Dalam proses pembelajaran, perangkat tes merupakan salah satu alat ukur. Agar pengukuran dapat memberikan hasil seperti yang diharapkan, maka diperlukan karakteristik alat ukur yang tepat. Dalam proses pembelajaran, sebelum menggunakan alat ukur perlu dilakukan identifikasi tentang karakteristik peserta didik yang akan diukur, karakteristik materi pembelajaran, jenis tingkah laku yang akan diukur, dan prosedur yang akan digunakan untuk melakukan pengukuran.

Pada umumnya, berdasarkan objek yang diukur, orang membedakan pengukuran atas dua hal, yaitu pengukuran fisik dan pengukuran psikologik. Pengukuran tinggi badan dan berat badan adalah contoh pengukuran fisik, sedangkan pengukuran mengenai tingkat kecerdasan dan tingkat kestabilan emosi seseorang adalah contoh pengukuran psikologik.

Pengukuran psikologik menjadi bahan kajian menarik dan menjadi bagian baku kurikulum mahasiswa psikologi dan pendidikan setelah pada tahun 1904 Thorndike mempublikasikan bukunya yang berjudul *An Introduction to the Theory of Mental and Social Measurement* (Crocker & Algina, 1986: 10). Teori-teori yang ada di buku tersebut kemudian disempurnakan oleh pakar-pakar pengukuran. Kumpulan *body of knowledge* tersebut, yang kemudian populer dengan nama teori tes klasik (*classical test theory*), memberikan dasar teori untuk pengembangan tes kecerdasan, tes prestasi, tes kepribadian, dan tes psikologik yang lain.

PENILAIAN

Banyak para ahli mendefinisikan penilaian (asesmen) secara berbeda. Johnson & Johnson (2002: 6) mendefinisikan "*assessment is collecting information about the quality or quantity of a change in student, group, teacher, or administrator*". Johnson & Johnson memandang penilaian sebagai suatu usaha untuk mengumpulkan informasi mengenai kuantitas atau kualitas dari adanya suatu perubahan yang terjadi pada peserta didik, kelompok, pendidik, atau pelaksana pendidikan. Pada definisinya, Johnson & Johnson menekankan kepada adanya perubahan (*change*) dan kuantitas dan kualitas perubahan itu yang merupakan fokus penilaian.

Di sisi lain, AERA, APA, & NCME (1999: 172) mengatakan bahwa "*assessment is any sistematic method of obtaining information from test and other sources, used to draw inferences about characteristic of people objects, or programs*". Berdasarkan definisi ini, penilaian adalah cara sistematis untuk memperoleh informasi, yang informasi ini dapat diperoleh dari suatu tes atau sumber lain, untuk melakukan kesimpulan mengenai karakteristik orang, objek, atau program yang dinilai. Menurut AERA, APA, & NCME, asesmen dapat dilakukan melalui tes atau non-tes.

Khususnya di bidang pendidikan, Popham (2005: 3) mendefinisikan "*educational assessment is a formal attempt to determine the status of a student respect to educational variables of interest*". Popham mendefinisikan penilaian pendidikan sebagai sebuah usaha formal untuk menentukan kedudukan atau status peserta didik terkait dengan variabel pendidikan yang ditentukan.

Permendikbud Nomor 104 Tahun 2014 tentang Penilaian Hasil Belajar oleh Pendidik mendefinisikan penilaian sebagai berikut.

> Penilaian hasil belajar oleh pendidik adalah proses pengumpulan informasi/bukti tentang capaian pembelajaran peserta didik dalam kompetensi sikap spiritual dan sikap sosial, kompetensi pengetahuan, dan kompetensi keterampilan yang dilakukan secara terencana dan sistematis, selama dan setelah proses pembelajaran.

Di luar definisi-definisi tersebut, masih banyak definisi penilaian yang dikemukakan orang. Namun demikian, pada prinsipnya terdapat kesamaan pandang mengenai penilaian. Pertama, penilaian menyimpulkan mengenai karakteristik atau variabel yang dipilih. Kedua, kesimpulan dari kegiatan penilaian adalah pernyataan mengenai kualitas, kuantitas, atau kedudukan sesuatu yang dinilai. Ketiga, kegiatan penilaian dilaksanakan secara sistematis dan terencana.

Menurut Permendikbud Nomor 104 Tahun 2014, ada 9 prinsip penilaian yang adalah sebagai berikut.

1. Sahih, berarti penilaian didasarkan pada data yang mencerminkan kemampuan yang harus diukur.
2. Objektif, berarti penilaian didasarkan kepada prosedur dan kriteria yang jelas, tidak dipengaruhi subjektivitas penilai.
3. Adil, berarti penilaian tidak menguntungkan atau merugikan peserta didik karena kebutuhan khusus serta perbedaan latar belakang agama, suku, budaya, adat istiadat, status sosial ekonomi, dan gender.
4. Terpadu, berarti penilaian oleh pendidik merupakan salah satu komponen yang tak terpisahkan dari kegiatan pembelajaran.
5. Terbuka, berarti prosedur penilaian, kriteria penilaian, dan dasar pengambilan keputusan dapat diketahui oleh pihak yang berkepentingan.
6. Holistik dan berkesinambungan, berarti penilaian oleh pendidik mencakup semua aspek kompetensi dan dengan menggunakan berbagai teknik pendilaian yang sesuai dengan kompetensi yang harus dikuasai peserta didik
7. Sistematis, berarti penilaian dilakukan secara berencana dan bertahap dengan mengikuti langkah-langkah baku.
8. Akuntabel, berarti penilaian dapat dipertanggungjawabkan, baik dari segi teknik, prosedur, maupun hasilnya.
9. Edukatif, berarti penilaian dilakukan untuk kepentingan dan kemajuan peserta didik dalam belajar.

Popham (1995) mengatakan ada 4 tujuan penilaian, yaitu untuk: (1) mendiagnosis kekuatan dan kelemahan peserta didik, (2) memonitor kemajuan peserta didik, (3) memberikan nilai (*grade*) pada peserta didik, dan (4) menentukan efektivitas pembelajaran yang dilakukan pendidik. Senada dengan Popham, Johnson & Johnson (2002), merumuskan 3 tujuan penilaian, yaitu untuk: (1) mendiagnosis pengetahuan dan keterampilan peserta didik, (2) memonitor kemajuan peserta didik terkait dengan tujuan pembelajaran, dan (3) menyediakan data untuk memberikan nilai kepada peserta didik.

Terkait dengan diagnosis kekuatan dan kelemahan peserta didik, dengan penilaian diharapkan para pendidik dapat mempunyai pengetahuan mengenai kekuatan dan kelemahan peserta didik dalam berbagai aspek tujuan pembelajaran yang telah dirancangnya. Terkait dengan monitoring kemajuan peserta didik, dengan penilaian diharapkan pendidik dapat menentukan apakah pendidik telah mendapatkan kemajuan seperti yang diharapkan. Bila tidak terjadi kemajuan seperti yang diharapkan, pendidik diwajibkan untuk melakukan suatu upaya profesional agar diperoleh kemajuan seperti yang diharapkan. Terkait dengan pemberian nilai kepada peserta didik, dengan penilaian diharapkan pendidik dapat memberikan nilai sebagai status final kemampuan peserta didik di akhir satuan pembelajaran. Akhirnya, terkait dengan penentuan efektivitas pembelajaran, dengan penilaian, pendidik akan mengetahui apakah proses pembelajaran yang telah

dirancangnya berjalan efektif atau tidak. Jika sebagian besar peserta didik mendapat nilai jelek pada akhir satuan pembelajaran, pada hal seharusnya tidak demikian, maka pembelajaran yang telah dilaluinya tidak dapat dikatakan efektif.

Jonhson & Johnson (2002) menggolongkan penilaian ke dalam tiga jenis, yaitu: penilaian diagnostik, penilaian formatif, dan penilaian sumatif. Penilaian diagnostik dilakukan untuk mengetahui kekuatan dan kelemahan peserta didik. Dengan penilaian diagnostik, para pendidik diharapkan dapat mengetahui kesalahan dan/atau miskonsepsi yang terjadi sebelum atau sesudah pembelajaran berlangsung. Penilaian ini dapat pula dipakai untuk mengumpulkan informasi mengenai apa yang telah diketahui dan yang belum diketahui oleh peserta didik.

Penilaian formatif dilaksanakan secara kontinu sepanjang satuan pembelajaran dengan tujuan utama untuk memperoleh balikan. Penilaian formatif merupakan bagian integral dari proses pembelajaran dengan dua alasan. Pertama, penilaian formatif memberikan balikan kepada peserta didik yang terkait dengan kemajuan yang telah ia capai. Kedua, penilaian formatif memberikan balikan kepada pendidik terkait dengan kemajuan proses pembelajaran yang dirancangnya dalam kaitannya dengan efektivitas pembelajaran yang menjadi tujuannya. Dengan penilaian formatif, kesalahan dan/atau miskonsepsi yang terjadi selama pembelajaran dapat dideteksi dan dicarikan jalan untuk memperbaikinya.

Penilaian sumatif mempunyai tujuan utama untuk menentukan kedudukan peserta didik terkait dengan tujuan pembelajaran yang telah dirancang. Dalam bahasa sederahana, penilaian sumatif mempunyai tujuan utama untuk memberikan nilai (*grade*) kepada peserta didik. Biasanya, penilaian sumatif dilakukan pada akhir satuan pembelajaran untuk menentukan status final peserta didik dalam kaitannya dengan tujuan pembelajaran yang telah dirancang oleh pendidik, Penilaian sumatif biasanya berbentuk ujian semester atau ujian akhir satuan pendidikan.

Untuk menentukan keberhasilan peserta didik dalam mengikuti proses pembelajaran, ada dua cara, yaitu penentuan keberhasilan berdasar-kan acuan norma (*norm-referenced*), yang sering disingkat PAN (Penilaian Acuan Norma), dan penentuan keberhasilan berdasarkan kriteria atau patokan (*criterion-referenced*), yang sering disingkat PAP (Penilaian Acuan Patokan).

Keberhasilan seorang peserta didik pada penilaian berdasar PAN dibandingkan dengan keberhasilan teman-teman sekelompoknya. Keberhasilan seorang peserta didik pada penilaian berdasar PAP dibandingkan dengan kriteria atau standar yang telah ditetapkan oleh pendidik sebelum pembelajaran pada satuan waktu pembelajaran berlangsung. Pelaksanaan penilaian berdasar PAP lebih kompleks daripada pelaksanaan penilaian

berdasar PAN. Pada pelaksanaan penilaian berdasar PAP, (1) *the domain of learning tasks be clearly defined,* (2) *the standards of performance be clearly specified and justified,* dan (3) *the measures of student achievement be criterion referenced* (Gronlund, 1985).

Pada umumnya, seorang pendidik tidak saja harus melakukan penilaian pada aspek kognitif, tetapi juga pada aspek afektif dan psikomotor. Dengan demikian, terdapat target penilaian untuk aspek kognitif, target penilaian untuk aspek afektif, dan target penilaian untuk aspek psikomotor.

Menurut Popham (1995), target penilaian aspek kognitif menitikberatkan kepada operasi intelektual (*intellectual operations*) peserta didik, target penilaian aspek afektif menitikberatkan kepada sikap (*attitudes*) dan nilai-nilai (*values*) yang dipunyai oleh peserta didik, dan target penilaian aspek psikomotor menitikberatkan kepada keterampilan gerak otot (*large-muscle and small-muscle skills*).

Senada dengan Popham, Anderson (1981) mengatakan bahwa aspek kognitif menitikberatkan kepada hal-hal yang berkaitan dengan cara berpikir (*typical ways of thinking*), aspek afektif menitikberatkan kepada hal-hal yang berkaitan dengan perasaan (*typical ways of feeling*), dan aspek psikomotor menitikberatkan kepada hal-hal yang berkaitan dengan cara tindak (*typical ways of acting*).[2]

ASUMSI-ASUMSI PADA PENILAIAN PENDIDIKAN

Menurut Reynolds, Livingston, dan Willson (2010: 9-13), ada beberapa asumsi yang melandasai penilaian pendidikan (*educational assessment*). Asumsi-asumsi itu secara ringkas disebutkan berikut ini.

1. *Psychological and educational construct exists.* Pada penilaian pendidikan, didefinisikan apa yang disebut konstruks. AERA, APA, dan NCME (1999) mendefinisikan konstruks sebagai kemampuan atau karakteristik yang diukur oleh suatu tes. Misalnya, prestasi belajar adalah suatu konstruks yang menyatakan pengetahuan atau pemahaman (*accomplishments*) seseorang pada suatu bidang yang telah diterimanya melalui pembelajaran[3]. Contoh lain konstruks adalah inteligensi dan sikap terhadap pembelajaran. Diasumsikan bahwa konstruks-konstruks tersebut ada.

---

[2] Pada Kurikulum 2013, didefinisikan adanya tiga domain tujuan pembelajaran, yaitu tujuan pembelajaran di domain pengetahuan, sikap, dan keterampilan. Perlu dilakukan pengkajian lebih lanjut, apakah tujuan di domain pengetahuan identik dengan tujuan di domain kognitif, apakah tujuan di domain sikap identik dengan tujuan di domain afektif, dan apakah tujuan di domain keterampilan identik dengan tujuan di domain psikomotor.

[3] Berdasarkan definisi ini, peserta didik yang dikenai tes prestasi belajar, harus sudah menerima pembelajaran terkait dengan materi tes. Tes potensi akademik, misalnya, bukanlah tes hasil belajar, yang berarti tidak diperlukan proses pembelajaran, ketika seseorang akan menempuh tes potensi akademik.

2. *Psychological and educational construct can be measured.* Cronbach (Reynolds, Livingston, dan William (2010: 10) mengemukakan adagium terkenal yang sering dikutip oleh para penganut pengukuran, yaitu "*If a thing exists, it exists in some amount. If it exists in some amount, it can be measured*[4]". Jadi, kalau konstruks itu ada, maka konstruks itu dapat diukur.
3. *Although we can measure construct, our measurement is not perfect.* Asumsi ini mengatakan bahwa walaupun konstruks dapat diukur, tetapi tidak pernah ada pengukuran yang sempurna (proses dan produknya). Oleh karena itu, diasumsikan bahwa ada *error* (kesalahan) pengukuran, walaupun mungkin kecil. Dengan asumsi inilah, para ahli terus menerus berusaha mengembangkan teori dan praksis pengukuran untuk memperkecil kesalahan pengukuran.
4. *There are different ways to measure any given construct.* Asumsi ini mengatakan bahwa suatu konstruks tertentu dapat diukur melalui berbagai macam cara, yang masing-masing cara mempunyai karakteristik sendiri-sendiri. Tes prestasi belajar, misalnya, dapat diukur dengan tes uraian dan dapat diukur pula dengan tes pilihan ganda, masing-masing cara mempunyai keunggulan dan kelemahannya sendiri-sendiri.
5. *All assessment procedures have strengths and limitations.* Walaupun suatu konstruks dapat diukur dengan berbagai macam cara, masing-masing cara itu mempunyai keunggulan dan kelemahan sendiri-sendiri. Ini berarti tidak ada suatu cara yang selalu baik untuk berbagai keadaan dan situasi[5].
6. *Multiple sources of information should be part of assessment process.* Asumsi ini mengatakan bahwa untuk menilai seseorang, harus digunakan berbagai sumber informasi. Ini akibat asumsi kelima yang mengatakan tidak ada satupun prosedur penilaian yang sempurna.
7. *Performance on tests can be generalized to nontest behaviors.* Diasumsikan bahwa segala sesuatu yang ada pada tes, misalnya cara pengembangannya, dapat dialihkan ke non-tes. Berdasar asumsi inilah

---

[4] Menurut Cronbach, jika sesuatu itu dibicarakan orang, berarti sesuatu itu ada. Jika sesuatu itu ada, maka sesuatu itu bisa diukur. Misalnya, orang sering membicarakan cinta. Menurut Cronbach, pasti keberadaan cinta itu dapat diukur. Kalaupun sampai sekarang, orang belum dapat mengukur cinta, misalnya, itu bukan berarti bahwa cinta tidak bisa diukur. Yang terja-di adalah belum dapat dibuat alat ukurnya. Oleh karena itu lah para psikolog berusaha terus menerus untuk menciptakan alat ukur mengenai sesuatu, walaupun oleh sebagian orang dianggap mustahil mengukur sesuatu itu.

[5] Walaupun diakui bahwa tes uraian lebih unggul dibandingkan dengan tes pilihan ganda dalam mengorganisir jawaban, tetapi Ujian Nasional, TIMSS, PISA, PIRLS, dan semacamnya akan selalu menggunakan tes pilihan ganda, karea jenis itulah yang paling cocok untuk melakukan pengujian pada skala besar. Penganjur agar Ujian Nasional menggunakan tes uraian tidaklah bijak, karena Ujian Nasional adalah ujian skala besar yang harus segera diumumkan hasilnya.

para ahli mengembangkan non-tes berdasarkan cara-cara yang dilakukan ketika para ahli tersebut mengembangkan tes.

8. *Assessment can provide information that helps educators make better educational decisions*. Penggunaan asesmen dalam pembelajaran diyakini dapat membantu pendidik untuk memperbaiki kinerjanya dalam proses pembelajaran. Berdasar asumsi ini, para praktisi pengukuran dan pengujian selalu berusaha menciptakan asesmen yang dapat membantu memperbaiki kualitas proses pembelajaran[6].
9. *Assessment can be conducted in a fair manner*. Diasumsikan bahwa penilaian dapat dilakukan dalam keadaan yang adil. Berdasarkan asumsi ini, para ahli pengukuran dan pengujian terus berusaha untuk membuat suatu prosedur agar pelaksanaan pengukuran dan pengujian berlangsung secara adil, tidak merugikan peserta didik yang berasal dari daerah atau etnis tertentu, misalnya.
10. *Testing and assessment can benefit our educational institutions and society as a whole*. Pada akhirnya, diasumsikan bahwa apa yang dilakukan oleh para ahli dan praktisi pengukuran dan pengujian diyakini akan berdampak positif terhadap lembaga-lembaga pendidikan dan masyarakat pendidikan secara keseluruhan.

Terkait dengan tugas pendidik (guru dan dosen), Reynolds, Livingston, dan Willson (2010: 25) mengatakan bahwa pendidik profesional haruslah dapat: (1) memilih dengan baik prosedur penilaian yang cocok untuk membuat keputusan pembelajaran (*instructional decision*), (2) mengembangkan dengan baik prosedur penilaian yang cocok untuk membuat keputusan pembelajaran, (3) melaksanakan penilaian, melakukan penskoran, dan menginterpretasi secara profesional penilaian yang dibuatnya, (4) menggunakan hasil penilaian dalam membuat keputusan pembelajaran, (5) mengembangkan prosedur pemberian skor (nilai) yang benar sesuai dengan informasi yang diperoleh dari penilaian, (6) mengkomunikasikan hasil penilaian kepada pihak-pihak terkait, dan (7) mengetahui dan menghindari tindakan tercela akibat penggunaan prosedur atau informasi penilaian yang tidak etis, *illegal*, dan tidak benar.

BAHAN DISKUSI

1. Ada yang menganggap bahwa penilaian dan pembelajaran adalah dua hal yang terpisah. Di sisi lain, ada yang mengganggap bahwa seharusnya pembelajaran da penilaian adalah dua kegiatan yang menyatu, tidak terpisahkan.

---

[6] Dikenal jenis penilaian yang disebut *assessment for learning* (penilaian untuk pembelajaran) yang tujuan utamanya memberikan balikan kepada peserta didik mengenai kesalahan-kesalahan yang diperbuatnya dalam mengerjakan soal-soal penilaian.

a. Berilah contoh praksis pembelajaran yang memberikan indikasi bahwa pembelajaran dan penilaian adalah dua kegiatan yang terpisah.
b. Berilah contoh praksis pembelajaran yang memberikan indikasi bahwa pembelajaran dan penilaian adalah dua kegiatan yang terintegrasi.
c. Perhatikan RPP (Rencana Pelaksanaan Pembelajaran) yang pernah Anda buat. Ketika Anda membuat RPP tersebut, paradigma manakah yang Anda pakai. pembejalaran dan penilaian merupakan dua kegiatan yang terpisah atau dua kegiatan yang menyatu? Mengapa?
c. Menurut Anda, manakah yang seharusnya dilakukan oleh pendidik. memandang pembalajaran dan penilaian sebagai dua kegiatan yang terpisah atau dua kegiatan yang menyatu? Mengapa?
d. Menurut Anda, manakah yang lebih meringankan tugas pendidik (guru, dosen), melaksanakan pembelajaran dengan paradigma pembelajaran dan penilaian merupakan dua kegiatan yang terpisah atau yang mempunyai paradigma yang mengatakan bahwa pembelajaran dan penilaian adalah dua kegiatan yang menyatu? Mengapa?
e. Menurut Anda, manakah yang lebih menguntungkan peserta didik (siswa, mahaasiswa) diberi pembelajaran oleh pendidik yang mempunyai paradigma pembelajaran dan penilaian merupakan dua kegiatan yang terpisah atau yang mempunyai paradigma bahwa pembelajaran dan penilaian adalah dua dua kegiatan yang menyatu? Mengapa?

2. Termasuk pengukuran fisik atau pengukuran psikologik, pengukuran mengenai hal berikut:

a. tinggi badan	f. tingkat kecerdasan
b. berat badan	g. motivasi
c. jarak tempuh	h. kedisiplinan belajar
d. kecepatan	i. ketekunan belajar
e. percepatan	j. penghargaan (*values*) terhadap matematika

3. Misalnya seseorang melakukan pengukuran mengenai kesetiaan pacarnya terhadap dirinya. Untuk itu, ia bersemedi, menerawang dengan kekuatan indera keenamnya, dan ia sampai kepada kesimpulan bahwa kesetiaan pacarnya terhadap dirinya hanya 40% saja (atau kesetiaan pacarnya hanya bernilai 40 dengan skala 100). Apakah orang tersebut melakukan pengukuran mengenai kesetiaan pacarnya berdasarkan konsep pengukuran yang dibicarakan di buku ini? Mengapa?

4. Misalnya kita percaya bahwa setiap hasil pengukuran selalu memuat kesalahan (*error*). Misalnya hasil pengukuran mengenai IQ Anda

dengan menggunakan alat ukur dan cara tertentu adalah 130. Skor 130 ini disebut skor tampak (*observed score*), dan dilambangkan dengan **X**. Andaikan IQ Anda sebenarnya adalah 140. Skor 140 ini disebut skor sebenarnya (*true score*) dan dilambangkan dengan **T**. Misalnya kesalahan pengukuran (*measurement error*) dilambangkan dengan **e**[7].

a. Menurut Anda, bagaimana hubungan (relasi) antara **X**, **T**, dan **e**?
b. Pada relasi yang Anda tuliskan, dapatkah **e** bernilai negatif? Nol? Positif?
c. Pada suatu pengukuran, dapatkah Anda memperoleh skor **T** (skor sebenarnya)? Mengapa?

5. Misalnya Anda melakukan pengukuran IQ seratus orang dengan alat ukur yang sama. Maka terdapat 100 skor tampak. $\mathbf{X}_1$, X_2, ... , $\mathbf{X}_{100}$., terdapat 100 skor sebenarnya, $\mathbf{T}_1$, T_2, ... , $\mathbf{T}_{100}$, terdapat 100 kesalahan pengukuran, $\mathbf{e}_1$, e_2, ... , $\mathbf{e}_{100}$.
   a. Menurut Anda, dapatkah semua **e** bernilai positif? Mengapa?
   b. Menurut Anda, dapatkah semua **e** bernilai negatif? Mengapa?
   c. Menurut Anda, dapatkah semua **e** bernilai nol? Mengapa?

6. Misalnya ada orang yang dapat melakukan pengukuran mengenai tinggi badan dan tingkat kecerdasan Anda dengan menggunakan alat ukur dan cara tertentu. Misalnya seseorang tersebut mengatakan kepada Anda bahwa menurut hasil pengukurannya, tinggi badan Anda adalah sekian cm dan tingkat kecerdasan Anda adalah sekian.
   a. Apakah Anda percaya benar hasil pengukuran tinggi badan Anda? Mengapa?
   b. Apakah Anda percaya benar hasil pengukuran tingkat kecerdasan Anda? Mengapa?
   c. Manakah yang lebih Anda percaya, hasil pengukuran tinggi badan Anda atau hasil pengukuran tingkat kecerdasan Anda? Memgapa?
   d. Jika Anda percaya bahwa pada setiap hasil pengukuran selalu memuat kesalahan (*error*) pengukuran, manakah yang kira-kira lebih besar kesalahan pengukurannya, kesalahan pengukuran pada pengukuran tinggi badan atau kesalahan pada pengukuran tingkat kecerdasan Anda? Mengapa?

7. Pada suatu pengukuran, manakah yang lebih disukai, pengukuran dengan kesalahan pengukuran yang kecil atau pengukuran dengan kesalahan pengukuran yang besar? Mengapa?

8. Pada suatu kelas terdapat 32 siswa. Diadakan ujian Matematika pada kelas tersebut. Dari skor ujian Matematika tersebut dicari rerata (μ)

---

[7] Dalam beberapa buku, kesalahan pengukuran dilambangkan dengan E.

dan deviasi baku-(σ)-nya. Diperoleh $\mu = 60$ dan $\sigma = 5$. Untuk memberi nilai siswa tersebut dalam skala lima (yaitu nilai dalam bentuk A, B, C, D, dan E) diberlakukan aturan konversi sebagai berikut.

Rentang Skor (X)	Nilai dalam Skala Lima
$X > \mu + 1{,}5\sigma$	A
$\mu + 0{,}5\sigma \leq X < \mu + 1.5\sigma$	B
$\mu - 0{,}5\sigma \leq X < \mu + 0{,}5\sigma$	C
$\mu - 1.5\sigma \leq X < \mu - 0.5\sigma$	D
$X < \mu - 1.5\sigma$	E

Jika aturan transformasi skornya seperti itu, penilaian tersebut menggunakan pendekatan PAN atau PAP? Mengapa?

9. Jika Anda seorang guru, akan menggunakan PAP atau PAN penilaian Anda?

10. Kurikulum Tingkat Satuan Pendidikan (KTSP) menggunakan PAP atau PAN?

11. Kurikulum 2013 menggunakan PAP atau PAN?

12. Menurut Anda, apa kelebihan dan kelemahan penilaian berdasar PAP?

13. Menurut Anda, apa kelebihan da kelemahan penilaian berdasar PAN?

14. Jika skor-skor yang diperoleh berdistribusi normal, aturan transformasi skornya seperti pada soal Nomor 8, berapa persenkah siswa yang memperoleh nilai A? Nilai B? Nilai C? Nilai D? Nilai E?

15. Misalnya skor ujian Matematika dari 32 siswa adalah sebagai berikut.

32	45	36	76	74	45	65	32
68	80	76	91	34	36	65	76
77	46	56	78	45	65	72	87
66	46	78	95	77	64	80	75

Pada kelas tersebut Siti mendapat skor 32 dan Amir mendapat skor 80. Dengan menggunakan transformasi seperti pada soal Nomor 8, berapakah nilai Siti dan berapakah nilai Amir?

16. Pada suatu kelas terdapat 32 siswa. Diadakan ujian Matematika pada kelas tersebut. Untuk memberi nilai siswa tersebut dalam skala lima diberlakukan aturan konversi sebagai berikut.

Rentang Skor (X)	Nilai dalam Skala Lima
$X > 80$	A
$70 \leq X < 80$	B
$60 \leq X < 70$	C
$50 \leq X < 60$	D
$X < 50$	E

Jika aturan transformasi skornya seperti itu, penilaian tersebut menggunakan pendekatan PAN atau PAP? Mengapa?

17. Jika diberlakukan aturan konversi skor seperti pada Soal Nomor 16, berapakah nilai Siti dan Amir pada Soal Nomor 13?

18. Seorang peneliti melakukan pengukuran mengenai motivasi siswa dengan menggunakan skala Likert. Ada 20 butir yang dipakai untuk melakukan pengukuran, masing-masing dengan alternatif jawaban SS (sangat setuju), S (setuju), N (netral), TS (tidak setuju) dan STS (sangat tidak setuju). Dari skor motivasi tersebut dicari rerata (μ) dan deviasi baku-(σ)-nya. Diperoleh $\mu = 60$ dan $\sigma = 15$. Peneliti mengelompokkan motivasi siswa ke dalam tiga kategori, yaitu Tinggi (T), Sedang (S), dan Rendah (R). Aturan pengelompokannya adalah sebagai berikut.

Rentang Skor (X)	Motivasi Siswa
$X > \mu + 0{,}5\sigma$	Tinggi
$\mu - 0{,}5\sigma \leq X \leq \mu + 0{,}5\sigma$	Sedang
$X < \mu - 0{,}5\sigma$	Rendah

Jika aturan transformasi skornya seperti itu, penilaian tersebut menggunakan pendekatan PAN atau PAP? Mengapa?

19. Seorang peneliti melakukan pengukuran mengenai motivasi siswa dengan menggunakan skala Likert. Ada 20 butir yang dipakai untuk melakukan pengukuran. Skor minimal yang mungkin adalah 0 dan skor maksimum yang mungkin adalah 100. Peneliti mengelompokkan motivasi siswa ke dalam tiga kategori, yaitu Tinggi (T), Sedang (S), dan Rendah (R). Aturan transformasinya adalah sebagai berikut.

Rentang Skor (X)	Motivasi Siswa
$X > 75$	Tinggi
$25 \leq X \leq 75$	Sedang
$X < 25$	Rendah

Jika aturan transformasi skornya seperti itu, penilaian tersebut menggunakan pendekatan PAN atau PAP? Mengapa?

20. Jika Anda seorang peneliti, aturan mana yang Anda pilih, seperti pada Soal Nomor 18 atau seperti pada Soal Nomor 19? Mengapa?

21. Misalnya Anda memilih aturan seperti pada Soal Nomor 18.
    a. Apakah pasti ada siswa dengan kategori motivasi Tinggi? Mengapa?
    b. Jika distribusi skornya normal, berapa persen siswa yang mempunyai motivasi Tinggi? Sedang? Rendah?[8]

22. Misalnya Anda memilih aturan seperti pada Soal Nomor 19. Apakah pasti ada siswa dengan kategori motivasi tinggi? Mengapa?

******

[8] Gunakan tabel distribusi normal baku yang ada pada kebanyakan buku-buku statistik.

BAB II
TEORI PENGUKURAN

PENDAHULUAN

Dewasa ini terdapat dua jenis teori pengukuran, yaitu teori tes klasik (*classical test theory*) dan teori tes modern yang lebih dikenal dengan teori respon butir (*item response theory*). Pada bab ini diperkenalkan secara sederhana kedua teori tersebut.

TEORI TES KLASIK

Pada teori tes klasik, terdapat 5 asumsi, yaitu sebagai berikut (Allen & Yen, 1979: 57).

1. $X = T + e$
2. $E(X) = T$
3. $\rho_{eT} = 0$
4. $\rho_{e_1 e_2} = 0$
5. $\rho_{e_1 T_2} = 0$

Asumsi-asumsi tersebut di atas dapat dinyatakan secara verbal sebagai berikut.

Asumsi Pertama: $X = T + e$

Pada model ini, skor yang diperoleh peserta tes, yang disebut skor amatan (*observed score*) X terdiri dari skor sebenarnya (*true score*) T dan kesalahan pengukuran e (*error score* atau *error of measurement*), yang dihubungkan oleh relasi $X = T + e$. Pada asumsi ini, yang diperoleh pada pengukuran adalah skor X, sedangkan skor T dan kesalahan e tidak dike-

tahui. Misalnya, pada pengukuran IQ. Amir mendapat skor 130, sehingga lalu dikatakan bahwa IQ Amir adalah 130. Skor 130 ini disebut skor tampak. Skor yang sesungguhnya (dalam arti IQ Amir yang sesungguhnya) tidak ada seorangpun yang tahu. Bisa jadi IQ Amir yang se-benarnya adalah 132 atau barangkali mungkin IQ Amir hanya 125. Jika IQ yang sebenarnya 132, maka kesalahan pengukurannya $e = -2$, sedangkan jika IQ yang sesungguhnya 125, maka kesalahan pengukurannya $e = 5$.

Ada dua macam kesalahan. yaitu kesalahan acak (*random error*) dan kesalahan sistematik (*systematic error*). Misalnya seseorang melakukan suatu pengukuran terhadap 100 orang. Berdasarkan teori pengukuran, maka terdapat 100 buah X. 100 buah T. dan 100 buah e seperti yang tampak pada Tabel 2.1.

Tabel 2.1 Skor Pengukuran terhadap 100 Orang

X	T	e	Relasi
X_1	T_1	e_1	$X_1 = T_1 + e_1$
X_2	T_2	e_2	$X_2 = T_2 + e_2$
X_3	T_3	e_3	$X_3 = T_3 + e_3$
X_4	T_4	e_4	$X_4 = T_4 + e_4$
X_5	T_5	e_5	$X_5 = T_5 + e_5$
...	...	...	...
X_{100}	T_{100}	e_{100}	$X_{100} = T_{100} + e_{100}$

Jika semua e adalah positif atau semua e adalah negatif, maka kesalahanya disebut kesalahan sistematik. Pada teori pengukuran, kesalahan yang terjadi diasumsikan merupakan kesalahan random (dalam arti ada e yang positif, ada e yang nol. dan ada e yang negatif).

Asumsi Kedua: $E(X) = T$

Di statistik matematik dikenal adanya nilai harapan dari suatu variabel random X yang dilambangi dengan $E(X)$. Nilai harapan jni merupakan rerata variabel random X pada populasinya, dan sering dilambangkan dengan $\mu = E(X)$. Asumsi $E(X) = T$ diartikan bahwa jika dilakukan pengukuran kepada orang yang sama dilakukan tak berhingga kali, sedangkan kemampuan orang tersebut sama dari satu pengukuran ke pengukuran yang lainnya, maka skor yang sesungguhnya T dapat dicari dengan mengambil rerata dari skor pengamatan X yang diperolehnya.

Implikasi praktis dari asumsi ini adalah bahwa untuk memberikan penilaian kepada seorang peserta didik, berilah ujian beberapa kali, mi-

salnya 4 kali dalam satu semester, kemudian dicari rerata dari 4 skor yang diperoleh. Rerata skor dari keempat skor merupakan skor final peserta didik tersebut dan dianggap merupakan skor yang sebenarnya. Semakin banyak dilakukan pengukuran, rerata skor tampaknya akan semakin mendekati skor sebenarnya.

Asumsi Ketiga $\rho_{eT} = 0$

ρ adalah lambang koefisien korelasi. Asumsi ketiga mengatakan bahwa pada sejumlah pengukuran, tidak ada korelasi antara kesalahan skor dan skor sebenarnya. Artinya, jika diperoleh skor sebenarnya T yang tinggi, kesalahan skornya e tidak harus tinggi dan sebaliknya jika diperolah skor sebenarnya T yang rendah, kesalahan skornya e tidak harus rendah. Pada konteks Tabel 2.1, maka diasumsikan tidak ada korelasi antara skor T dan skor e pada Tabel 2.2.

Tabel 2.2. Skor T dan e pada 100 Kali Pengukuran

T	e
T_1	e_1
T_2	e_2
T_3	e_3
T_4	e_4
T_5	e_5
...	...
T_{100}	e_{100}

Asumsi Keempat $\rho_{e_1e_2} = 0$

Misalnya dilakukan pengukuran kepada 100 orang dengan menggunakan dua tes yaitu Tes A dan Tes B. Masing-masing tes menghasilkan skor tampak, skor sebenarnya, dan kesalahannya masing-masing. Misalnya hasil pengukuran pada Tes A (tes pertama) dan pada Tes B (tes kedua) tampak pada Tabel 2.3.

Asumsi keempat mengatakan bahwa tidak ada korelasi antara e_1 dan e_2. Artinya jika kesalahan pengukuran pertama pada Tes A tinggi, kesalahan pengukuran pertama pada Tes B tidak harus tinggi; sebaliknya jika kesalahan pengukuran pertama pada Tes A rendah, kesalahan pengukuran pertama pada tes B tidak harus rendah. Demikian pula untuk kesalahan kedua, kesalahan ketiga, dan seterusnya.

Tabel 2.3. Skor Pengukuran terhadap 100 Orang dengan Menggunakan Tes A dan Tes B

Hasil Pengukuran pada Tes Pertama (Tes A)			Hasil Pengukuran pada Tes Kedua (Tes B)		
X_1	T_1	e_1	X_2	T_2	e_2
X_{11}	T_{11}	e_{11}	X_{12}	T_{12}	e_{12}
X_{21}	T_{21}	e_{21}	X_{22}	T_{22}	e_{22}
X_{31}	T_{31}	e_{31}	X_{32}	T_{32}	e_{32}
X_{41}	T_{41}	e_{41}	X_{42}	T_{42}	e_{42}
X_{51}	T_{51}	e_{51}	X_{52}	T_{52}	e_{512}
...	...	...	...	...	...
X_{1001}	T_{1001}	e_{1001}	X_{1002}	T_{1002}	e_{1002}

Asumsi Kelima: $\rho_{e_1T_2} = 0$

Misalnya dilakukan pengukuran kepada 100 orang dengan menggunakan dua tes yaitu Tes A dan Tes B. Seperti disebutkan di muka, masing-masing tes menghasilkan skor tampak, skor sebenarnya, dan kesalahannya masing-masing. Misalnya hasil pengukuran pada Tes A (tes pertama) dan pada Tes B (tes kedua) tampak pada Tabel 2.3.

Asumsi kelima mengatakan bahwa tidak ada korelasi antara e_1 dan T_2. Artinya jika kesalahan pengukuran pertama pada tes pertama tinggi, skor sebenarnya yang pertama pada tes kedua tidak harus tinggi. Sebaliknya, jika kesalahan pengukuran pertama pada tes pertama rendah. skor sebenarnya yang pertama pada tes kedua tidak harus rendah,

Kecuali kelima asumsi tersebut, didefinisikan adanya dua tes paralel (*parallel test*) dan dua tes ekuivalen-τ (*essentially* τ*-equivallent tests*) sebagai berikut (Allen & Yen, 1979: 57).

Dua Tes Paralel

Jika dua tes (yaitu tes pertama dan tes kedua) mempunyai skor tampak X dan X' yang memenuhi asumsi 1 s.d. 5. dan jika untuk setiap populasi peserta tes berlaku T = T' dan $\sigma_e^2 = \sigma_{e'}^2$. maka dua tes tersebut disebut tes paralel

Jika misalnya terdapat dua tes yaitu tes A dan tes B, Misalnya kedua tes tersebut dikenakan kepada 100 orang. Misalnya skor pengukurannya tampak pada Tabel 2.3. Jika misalnya $T_{i1} = T_{i2}$ untuk setiap i = 1, 2, ... ,

100 dan $\sigma^2_{e1} = \sigma^2_{e2}$. maka kedua tes disebut dua tes paralel. Dalam konteks ini, Tes B adalah paralelnya tes A, dan sebaliknya Tes A adalah paralelnya tes B.

Untuk selanjutnya diperjanjikan penggunaan notasi X' yang menyatakan paralelnya tes X. Jadi diperjanjikan tes X' adalah paralelnya tes X dan tes X adalah paralelnya tes X'. Konsep mengenai dua tes paralel menjadi sangat penting, terutama untuk mendefinisikan koefisien reliabilitas. Didefinisikan koefisien reliabilitas tes X adalah koefisien korelasi antara skor tampak tes X dan skor tampak tes X', dan dilambangkan dengan $\rho_{XX'}$. Pada sampel, koefisien reliabilitas dilambangkan dengan $r_{XX'}$. Untuk efisiensi, kadang koefisien reliabilitas pada sampel dilambangkan dengan r_{11}.

Dua Tes Ekuivalen-τ

Jika dua tes mempunyai skor tampak X_1 dan X_2 yang memenuhi asumsi 1 s.d. 5, dan jika untuk setiap populasi peserta tes berlaku $T_1 = T_2 + k$ dengan k konstanta, maka dua tes tersebut disebut tes τ-ekuivalen

Teorema-teorema pada Teori Pengkuran Klasik

Berdasarkan asumsi-asumsi 1 s.d. 5 di atas, dapat diturunkan sejumlah teorema. Allen & Yen (1979: 61-65) menurunkan 18 teorema dari kelima asumsi di muka. Namun pada buku ini dibahas beberapa saja yang penting.

Teorema 1

$E(e) = 0$

Teorema ini mengatakan bahwa jika kepada seseorang diberikan tes yang sama berulang-ulang, maka rerata kesalahan yang diperoleh adalah nol.

Bukti:

$X = T + e$	asumsi 1
$E(X) = E(T + e)$	asumsi 1
$E(X) = E(T) + E(e)$	sifat nilai harapan
$E(X) = T + E(e)$	sebab T suatu konstanta
$T = T + E(e)$	asumsi 2
$E(e) = 0$	terbukti

Teorema 2

$$\sigma_X^2 = \sigma_T^2 - \sigma_e^2 \qquad 2.1$$

Teorema ini mengatakan bahwa pada pengukuran kepada sejumlah orang (atau pengukuran kepada orang yang sama sebanyak n kali), maka variansi skor amatan sama dengan variansi skor sebenarnya ditambah dengan variansi kesalahannya. Untuk selanjutnya didefinisikan σ_e sebagai kesalahan baku pengukuran (*the standard of error measurement*).

Bukti:

$X = T + e$	asumsi 1
$\sigma_X^2 = \sigma_{T+e}^2$	asumsi 1
$\sigma_X^2 = \sigma_T^2 + \sigma_e^2 + 2\sigma_{Te}$	sifat vatiansi
$\sigma_X^2 = \sigma_T^2 + \sigma_e^2 + 0$	asumsi 2 (ingat $\rho_{Te} = \frac{\sigma_{Te}}{\sigma_T \sigma_e}$)
$\sigma_X^2 = \sigma_T^2 + \sigma_e^2$	terbukti

Berdasarkan Persamaan 2.1 pada Teorema 2, diperoleh hubungan berikut.

$$\sigma_T^2 \leq \sigma_X^2 \text{ dan } \sigma_e^2 \leq \sigma_X^2 \qquad 2.2$$

Teorema 3

$\sigma_X^2 = \sigma_{X'}^2$ (X dan X' adalah skor amatan dari dua tes paralel)

Teorema 3 menyatakan bahwa jika dua tes yang paralel dikenakan pada sekelompok orang, maka dua tes tersebut menghasilkan skor amatan yang sama pada masing-masing orang.

Bukti:

$\sigma_X^2 = \sigma_T^2 + \sigma_e^2$	Teorema 2
$= \sigma_{T'}^2 + \sigma_{e'}^2$	definisi tes paralel
$= \sigma_{X'}^2$	Teorema 2
	(terbukti)

Teorema 4

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} \qquad 2.3$$

Teorema 4 menyatakan bahwa koefisien korelasi antara skor amatan X dan skor amatan X' adalah rasio antara variansi skor sebenarnya terhadap variansi skor amatannya.

Bukti:

$$\rho_{XX'} = \frac{\sigma^2_{XX'}}{\sigma^2_X \sigma^2_{X'}} \qquad \text{definisi koefisien korelasi}$$

$$= \frac{\sigma_{(T+E)(T'+E')}}{\sigma_X \sigma_X} \qquad \text{asumsi 1, teorema 3}$$

$$= \frac{\sigma_{TT'} + \sigma_{TE'} + \sigma_{ET'} + \sigma_{EE'}}{\sigma^2_X} \qquad \text{sifat kovariansi}$$

$$= \frac{\sigma^2_T + 0 + 0 + 0}{\sigma^2_X} \qquad \text{asumsi 3, 4, dan 5}$$

$$= \frac{\sigma^2_T}{\sigma^2_X} \qquad \text{terbukti}$$

Seperti disebutkan di depan, $\rho_{XX'}$ merupakan koefisien reliabilitas tes X atau koefisien reliabilitas tes X'. Dengan memperhatikan bahwa $\rho_{XX'} = \frac{\sigma^2_T}{\sigma^2_X}$, variansi tidak pernah negatif, dan $\sigma^2_T \leq \sigma^2_X$, maka dapat disimpulkan bahwa $0 \leq \rho_{XX'} \leq 1$[1]. Ini berarti bahwa secara teoretis, rentang koefisien reliabilitas adalah antara 0 dan 1. Jika tes X memunyai $\rho_{XX'} = 0$ maka tes tersebut merupakan tes yang yang sama sekali tidak reliabel, sedangkan jika tes X mempunyai $\rho_{XX'} = 1$ maka tes tersebut merupakan tes yang reliabel sempurna.

Perlu diketahui bahwa pada kenyataannya, koefisien reliabitas suatu tes tidak dapat dihitung dengan menggunakan rumus $\rho_{XX'} = \frac{\sigma^2_T}{\sigma^2_X}$, karena pada suatu pengukuran, tidak pernah diperoleh skor sebenarnya (T). Yang diperoleh adalah skor tampak (X). Oleh karena itu, para pakar memper-

[1] Ini berarti kalau ada orang yang mengestimasi koefisien reliabilitas suatu instrumen dan diperoleh koefisien reliabilitasnya negatif atau lebih dari satu, maka perlu dicek ulang cara menghitungnya, karena secara teoretis koefisien reliabilitas tidak pernah negatif dan tidak pernah lebih dari satu.

kenalkan rumus-rumus untuk **mengestimasi** koefisien reliabilitas. Misalnya Kuder dan Richardson memperkenalkan rumus KR-20[2] dan KR-21 dan Cronbach memperkenalkan rumus alpha. Di luar rumus itu, masih banyak rumus yang diperkenalkan oleh para pakar pengukuran untuk mengestimasi koefisien reliabilitas.

Teorema 5

$$\rho_{XX'} = 1 - \frac{\sigma_e^2}{\sigma_X^2} \qquad 2.4$$

Formula ini adalah bentuk lain dari formula koefisien reliabilitas tes X.

Bukti:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} \qquad \text{teoema 4}$$

$$= \frac{\sigma_X^2 - \sigma_e^2}{\sigma_X^2} \qquad \text{Teorema 2}$$

$$= 1 - \frac{\sigma_e^2}{\sigma_X^2}$$

Tampak bahwa rumus pada Persamaan 2.4 adalah turunan langsung dari Persamaan 2.3.

Teorema 6
Jika X adalah jumlah skor dari N tes paralel Y_i demikian hingga $X = \sum Y$, maka $\sigma_{TX}^2 = N^2 \sigma_{TY}^2$

Pada kasus ini diasumsikan terdapat N tes paralel $Y_1, Y_2, Y_3, \ldots Y_N$ dan ke-N tes tersebut digabung menjadi satu tes X, yang dilambangkan dengan $X = \sum Y$, maka variansi skor sebenarnya dari tes X sama dengan N^2 kali variansi tes Y.

Bukti

(diserahkan kepada pembaca)

---

[2] Disebut rumus KR-20 karena rumus itu merupakan publikasi Kuder dan Richard-son yang ke-20, sedangkan KR-21 adalah publikasi mereka berdua yang ke-21

Teorema 7

$$\rho_{XX'} = \frac{N\rho_{YY'}}{1+(N-1)\rho_{YY'}} \quad \text{(rumus Spearman-Brown)} \qquad 2.5$$

Bukti:

(diserahkan kepada pembaca)

Rumus pada Persamaan 2.5 menunjukkan bahwa semakin panjang suatu tes, maka tes tersebut semakin reliabel. Pembaca dapat mencoba mensubstikan berbagai nilai N pada Persamaan 2.5 dan akan mendapatkan kenyataan bahwa semakin panjang suatu tes, maka akan semakin tinggi koefisien reliabilitasnya.

Contoh 2.1

Suatu tes dengan panjang 30 butir mempunyai koefisien reliabilitas 0.60. Jika tes tersebut diperpanjang menjadi 40 butir, berapa koefisien reliabilitas tes yang baru?

Jawab:

$$N = \frac{40}{30} = 1{,}33;\ \rho_{YY'} = 0{,}60;\ \rho_{XX'} = ?$$

$$\rho_{XX'} = \frac{N\rho_{YY'}}{1+(N-1)\rho_{YY'}} = \frac{(1.33)(0.60)}{1+(1{,}33-1)(0{,}60)} = \frac{0{,}798}{1{,}198} = 0{,}67$$

Jadi, koefisien reliabilitas tes yanng baru adalah 0,67

Teorema 8

Jika $\rho_{YY'} \neq 0$, maka $\lim\limits_{N\to\infty} \rho_{XX'} = 1$

Bukti:

(diserahkan kepada pembaca)

Teorema 8 menyatakan bahwa walaupun kalau tes diperpanjang koefisien reliabilitasnya meningkat, namun koefisien reliabilitas suatu tes tidak akan melebihi satu walaupun tes tersebut diperpanjang terus menerus. Hal ini sekaligus merupakan bukti bahwa nilai maksimum koefisien reliabilitas adalah 1.

TEORI RESPONS BUTIR

Menurut Dali S. Naga (1992: 4), pada pengukuran berdasar teori tes klasik, tes yang sama yang dijawab oleh kelompok peserta tes yang sama menghasilkan karakteristik yang sama pula, tetapi jika kelompok butir soal yang sama dijawab oleh kelompok peserta yang berbeda menghasilkan

karakteristik yang berbeda. Dengan kata lain, karakteristik butir soal dipengaruhi oleh peserta tes yang menempuh tes tersebut. Di sisi lain, jika kelompok peserta yang sama menempuh tes yang berbeda, maka ciri kelompok peserta itu pada umumnya berubah. Ini berarti, ciri-ciri kelompok peserta tes berubah jika mereka menempuh tes yang berbeda.

Untuk mengatasi kelemahan-kelemahan yang ada pada teori tes klasik, para ahli pengukuran berusaha mencari model alternatif. Model yang diinginkan harus mempunyai sifat-sifat: (1) karakteristik butir soal tidak tergantung kepada kelompok peserta tes yang dikenai butir soal tersebut, (2) skor yang menyatakan kemampuan peserta tes tidak tergantung kepada tes, (3) model dinyatakan dalam tingkatan (*level*) butir soal, tidak dalam tingkatan tes, (4) model tidak memerlukan tes paralel untuk menghitung koefisien reliabilitas, dan (5) model menyediakan ukuran yang tepat untuk setiap skor kemampuan (Hambleton, Swaminathan, & Rogers, 1991: 5). Model alternatif yang dapat mempunyai ciri-ciri itu adalah model pengukuran yang disebut teori respons butir (*item response theory*).

Model pengukuran pada teori respons butir berdasarkan dua postulat, yaitu: (1) kinerja peserta tes pada suatu butir soal dapat diprediksi oleh sekumpulan faktor yang disebut *traits* atau kemampuan (*abilities*), dan (2) hubungan antara kinerja peserta tes pada suatu butir soal dan sekumpulan *traits* dapat digambarkan dalam sebuah fungsi monoton naik yang disebut fungsi karakteristik butir (*item characteristic function*) atau kurva karakteristik butir (*item charateristic curve*) (Hambleton, Swaminathan, & Rogers, 1991: 7). Fungsi karakteristik butir ini menyatakan bahwa semakin meningkat level kemampuan seseorang, semakin meningkat pula peluangnya menjawab benar suatu butir tertentu. Namun demikian, peningkatan level kemampuan seseorang tidak berbanding lurus dengan peluangnya menjawab benar suatu butir tertentu.

Asumsi-asumsi pada Teori Respons Butir

Ada tiga asumsi dasar yang mendasari teori pengukuran berdasar teori repons butir, yaitu: (1) unidimensionalitas, (2) independensi lokal, dan (3) fungsi karakteristik butir menyatakan hubungan yang sebenarnya antara variabel yang tak terobservasi (yaitu kemampuan) dengan variabel terobservasi (yaitu respons butir) (Hambleton, Swaminathan, & Rogers, 1991: 9; Sumadi Suryabrata, 2000: 28). Asumsi unidimensionalitas dan independensi lokal dapat dijelaskan sebagai berikut.

Asumsi unidimensionalitas menyatakan bahwa hanya satu kemampuan yang diukur oleh sekumpulan butir-butir soal dalam suatu tes. Asumsi ini pada praktik sukar dipenuhi, sebab terdapat banyak faktor yang dapat mempengaruhi hasil suatu tes. Faktor-faktor tersebut antara lain tingkat motivasi, kecemasan, kemampuan untuk bekerja cepat, dan keterampilan

kognitif lain di luar kemampuan yang diukur oleh sekumpulan butir soal dalam suatu tes. Hal yang dimaksud dengan unidimensionalitas dalam hal ini adalah adanya faktor dominan yang mempengaruhi hasil suatu tes. Faktor dominan itulah yang disebut kemampuan yang diukur oleh suatu tes.

Asumsi independensi lokal menyatakan bahwa jika kemampuan yang mem-pengaruhi suatu tes adalah konstan, maka respons peserta tes pada setiap pasangan butir soal adalah independen secara statistik. Dengan kata lain, asumsi independensi lokal menyatakan bahwa tidak ada korelasi antara respons peserta tes pada butir soal yang berbeda. Hal ini juga berarti bahwa kemampuan yang dinyatakan dalam model adalah satu-satunya faktor yang mempengaruhi respons peserta tes pada butir-butir soal.

Model-model pada Teori Respons Butir Unidimensional

Ada tiga model yang populer pada teori respons butir, yang cocok untuk tes dikhotomous (termasuk tes pilihan ganda), yang disebut model logistik satu parameter, model logistik dua parameter, dan model logistik tiga parameter.

Model Logistik Satu Parameter

Model logistik satu parameter sering disebut juga dengan model Rasch, sebagai penghargaan kepada penemunya. Fungsi karakteristik butir untuk model logistik satu parameter ditentukan dengan persamaan (Hambleton, Swaminathan, & Rogers, 1991: 12):

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}; \; i = 1, 2, 3, \ldots, n \qquad 2.6$$

dengan

$P_i(\theta)$ = peluang seseorang dengan kemampuan θ menjawab butir soal ke-i dengan benar

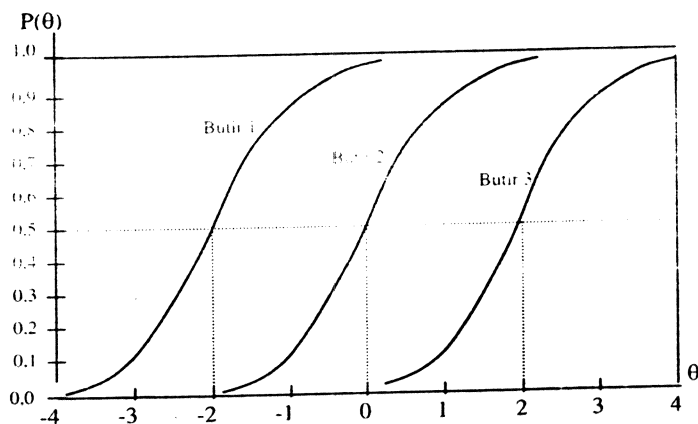
b_i = parameter tingkat kesulitan untuk butir soal ke-i

n = banyaknya butir soal dalam tes

e = bilangan pokok logaritma natural, yang nilainya adalah 2,718, jika dibulatkan ke tiga angka di belakang koma

Parameter tingkat kesulitan, yaitu b, untuk sebuah butir soal adalah titik pada skala kemampuan, yang pada titik itu peluang menjawab benar butir tersebut sebesar 0,5[3] (Hambleton, Swaminathan, & Rogers, 1991: 13).

[3] Perhatikanlah bahwa pendefinisian tingkat kesulitan butir pada Teori Respons Butir ini berbeda dengan pendefinisian tingkat kesulitan butir pada Teori Tes Klasik. Seperti diketahui, pada teori tes klasik, tingkat kesulitan suatu butir adalah proporsi peserta tes yang menjawab benar butir tersebut dengan banyaknya keseluruhan peserta tes.

Jika kemampuan (θ) ditransformasikan demikian hingga mempunyai rerata (*mean*) 0 dan simpangan baku 1, maka nilai b biasanya berkisar antara −2 sampai dengan 2 (Hambleton, Swaminathan, & Rogers, 1991: 13). Butir soal yang tingkat kesulitannya mendekati −2 merupakan butir soal yang sangat mudah dan butir soal yang tingkat kesulitannya mendekati 2 merupakan butir soal yang sangat sukar.

Gambar 2.1. Kurva Karakteristik Butir dengan Tingkat Kesulitan yang Berbeda

Pada Gambar 2.1, peserta tes dengan kemampuan θ = −2 mempunyai peluang sebesar 0,5 untuk menjawab benar butir soal nomor 1, peserta tes dengan kemampuan θ = 0 mempunyai peluang sebesar 0,5 untuk menjawab benar butir soal nomor 2, dan peserta tes dengan kemampuan θ = 2 mempunyai peluang sebesar 0,5 untuk menjawab benar butir soal nomor 3. Dengan demikian, pada Gambar 2.1, butir soal nomor 1 mempunyai tingkat kesulitan sebesar b = −2, butir soal nomor 2 mempunyai tingkat kesulitan b = 0, dan butir soal nomor 3 mempunyai tingkat kesulitan b = 2. Perhatikan bahwa kurva-kurva tersebut berbeda hanya pada letaknya saja. Kurva-kurva tersebut saling sejajar. Ini berarti hanya tingkat kesulitan butir saja yang mempengaruhi kinerja peserta tes. Pada model logistik satu paramer, daya pembeda masing-masing butir sama dan tidak ada unsur tebakan dalam menjawab butir soal.

Perhatikanlah bahwa kurva karakteristik butir Pada Gambar 2.1 berbentuk seperti huruf S, tidak berbentuk garis lurus. Asumsi yang melandasinya adalah bahwa hubungan antara kemampuan peserta tes dengan peluangnya menjawab benar butir tes tersebut fidak berbanding lurus.

Asumsi lain yang perlu diperhatikan bahwa Pada Gambar 2.1. kurvanya mempunyai asimtot di P = 1 dan P = 0. Ini berarti semakin pandai seseorang, maka semakin tinggi peluangnya menjawab benar suatu butir soal. Namun demikian, betapapun pandai seseorang, peluangnya menjawab benar butir tersebut tidak akan pernah sama dengan satu. Ini berarti, pada Teori Respons Butir, pada skala seratus, tidak ada seseorang yang mendapat nilai seratus. Sebaliknya, semakin bodoh seseorang, semakin kecil peluangnya menjawab benar suatu butir. Namun demikian, betapapun bodoh seseorang, peluang seseorang menjawab benar suatu butir tidak akan pernah nol. Ini berarti, pada Teori Respons Butir, tidak ada seseorang yang mendapat nilai 0. Hal ini berbeda dengan penskoran pada Teori Tes Klasik yang memungkinkan seseorang untuk mendapatkan nilai 0 atau mendapatkan nilai 100.

Grafik pada Gambar 2.1 menunjukkan bahwa kurvanya kontinu. Ini berarti pada Teori Respons Butir, skor-skor peserta tes bersifat kontinu. Di sisi lain, pada Teori Tes Klasik, skor-skor peserta tes bersifat deskrit. Pada Teori Tes Klasik, jika terdapat 20 butir soal, maka nilai peserta tes adalah 0, 5, 10, ... , 90, 95, 100 yang bersifat deskrit.

Pada Teori Respons Butir, parameter b_i diestimasi berdasarkan data empirik sebaran peserta tes dengan cara estimasi tertentu, misalnya dengan *maxsimum likelihood*. Ini hanya bisa dikerjakan oleh suatu program komputer tertentu, misalnya program komputer Bilog. Faktor inilah yang membuat Teori Respons Butir tidak mudah diimplementasikan.

Penskoran pada Teori Respons Butir dilakukan dengan memperhatikan sebaran jawaban peserta tes di mana tingkat kesulitan butir diperhatikan. Artinya, misalnya Amir menjawab benar 4 butir soal, yaitu butir soal nomor 1, 2 3, dan 4. Di sisi lain, Parti menjawab 4 butir soal, tetapi pada nomor lain, misalnya pada butir soal nomor 5, 6, 7, dan 8. Maka skor yang diperoleh Amir berbeda dengan skor yang diperoleh Parti, walaupun mereka sama-sama menjawab 4 butir soal, karena tingkat kesulitan ke-delapan soal itu tidaklah sama. Ini berbeda dengan penskoran pada Teori Tes Klasik, di mana penskoran tidak memperhatikan tingkat kesulitan masing-masing butir, semua butir dianggap mempunyai tingkat kesulitan yang sama. Yang berarti bahwa pada Teori Tes Klasiik, skor Amir dan skor Parti sama, karena mereka menjawab benar sama-sama 4 butir.

2) Model Logistik Dua Parameter

Pada tahun 1952, Lord mengembangkan model respons butir dua parameter dengan mendasarkan pada *ogive* distribusi normal. Lord dipandang sebagai orang pertama yang mengembangkan model respons butir dua parameter (Hambleton, Swaminathan, & Rogers, 1991: 14). Kemudian, pada tahun 1968, Birnbaum mengembangkannya menjadi model logistik

dua parameter dengan persamaan berikut (Hambleton, Swaminathan, & Rogers, 1991: 15):

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \; ; i = 1, 2, 3, \ldots, n \qquad 2.7$$

dengan

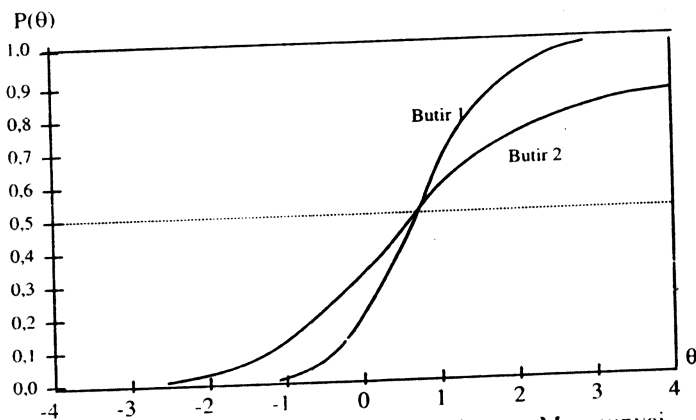
$P_i(\theta)$ = peluang seseorang dengan kemampuan θ menjawab butir soal ke-i dengan benar;

a_i = parameter daya pembeda untuk butir soal ke-i

b_i = parameter tingkat kesulitan untuk butir soal ke-i

n = banyaknya butir soal dalam tes

e = bilangan pokok logaritma natural, yang nilainya adalah 2,718, jika dibulatkan ke tiga angka di belakang koma

D = faktor skala, yang nilainya diambil sebesar 1,7, yaitu simpangan baku distribusi logistik yang paling mendekati distribusi normal.

Gambar 2.2. Kurva Karakteristik Butir yang Mempunyai Tingkat Kesulitan Sama tetapi Mempunyai Daya Pembeda yang Berbeda

Parameter daya pembeda, yaitu a, proporsional terhadap koefisien arah garis singgung (*slope*) pada titik θ = b (tingkat kesulitan)[4] (Hambleton, Swaminathan, & Rogers, 1991: 15). Butir soal yang mempunyai daya pembeda yang besar mempunyai kurva yang sangat menanjak, sedangkan

---

[4] Perhatikan bahwa definisi daya pembeda butir pada Teori Respons Butir juga berbeda dengan definisi daya pembeda pada Teori Tes Klasik. Pada Teori Tes Klasik, daya pembeda butir soal didefinisikan sebagai selisih antara proporsi kelompok atas yang menjawab benar butir tersebut dengan proporsi kelompok bawah yang menjawab benar butir tersebut.

butir soal yang mempunyai daya pembeda yang kecil mempunyai kurva yang sangat landai. Secara teoretis, daya pembeda dapat mempunyai nilai mulai dari $-\infty$ sampai dengan $+\infty$. Namun demikian, untuk butir soal yang baik, nilai parameter a harus terletak antara 0 dan 2 (Hambleton, Swaminathan, & Rogers, 1991: 15).

Dua butir soal pada Gambar 2.2 mempunyai tingkat kesulitan yang sama, namun mempunyai daya pembeda yang berbeda. Daya pembeda untuk butir soal nomor 1 lebih besar daripada daya pembeda untuk butir soal nomor 2. Berbeda dengan kurva-kurva pada model logistik satu parameter, kurva-kurva pada model logistik dua parameter tidak saling sejajar. Persamaan fungsi karakteristik butir pada persamaan (2), dapat ditulis dalam bentuk lain sebagai berikut:

$$P_i(\theta) = \frac{1}{1+e^{-Da_i(\theta-b_i)}} \; ; i = 1, 2, 3, \ldots, n \qquad 2.8$$

Persamaan 2.8 diperoleh dari Persamaan 2.7 dengan mengalikan pembilang dan penyebut ruas kanan Persamaan 2.7 dengan $e^{-Da_i(\theta-b_i)}$.

3) Model Logistik Tiga Parameter

Persamaan fungsi karakteristik butir untuk model logistik tiga parameter adalah sebagai berikut:

$$P_i(\theta) = c_i + (1-c_i)\frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}} \; ; i = 1, 2, 3, \ldots, n \qquad 2.9$$

dengan

$P_i(\theta)$ = peluang seseorang dengan kemampuan θ menjawab butir soal ke-i dengan benar

a_i = parameter daya pembeda untuk butir soal ke-i

b_i = parameter tingkat kesulitan untuk butir soal ke-i

c_i = parameter tebakan (*pseudo-guessing*) untuk butir soal ke-i

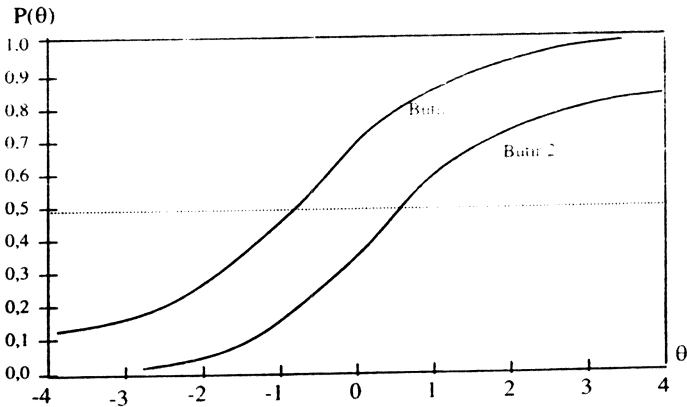
n = banyaknya butir soal dalam tes

e dan D sama seperti pada model logistik dua parameter

Model logistik tiga parameter memperbolehkan adanya asimtot bawah yang tidak nol, yang berarti model ini mengijinkan adanya faktor tebakan, seperti yang terjadi pada tes pilihan ganda. Dua butir soal pada Gambar 2.3 mempunyai daya pembeda yang sama namun mempunyai unsur tebakan yang berbeda. Butir soal nomor 1 mempunyai faktor tebakan

yang lebih besar (c = 0,1) dibandingkan faktor tebakan pada butir soal nomor 2 (c = 0). Persamaan (4) dapat ditulis sebagai berikut:

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-Da_i(\theta - b_i)}} \; ; \; i = 1, 2, 3, \ldots, n \qquad 2.10$$

Persamaan 2.10 diperoleh dari Persaman 2.9 dengan mengalikan pembilang dan penyebut suku kedua ruas kanan persamaan 2.9 dengan $e^{-Da_i(\theta - b_i)}$.

Gambar 2. 3. Kurva Karakteristik Butir yang Mempunyai Daya Pembeda Sama, tetapi Mempunyai Faktor Tebakan Berbeda

PERBANDINGAN TEORI TES KLASIK DAN TEORI RESPONS BUTIR

Teori tes klasik memuat berbagai keunggulan dan kelemahan. Keunggulan teori tes klasik, antara lain: (a) menggunakan konsep yang sederhana untuk menentukan kemampuan peserta tes, (b) menggunakan konsep yang sederhana dalam menghitung koefisien validitas dan reliabilitas tes serta menghitung nilai parameter butir soal, (c) dapat digunakan pada sampel kecil, misalnya pada tingkat kelas, (d) sudah digunakan dalam praksis pengukuran dan pengujian dalam kurun waktu yang lama, sehingga telah diketahui dan dipahami oleh sebagian besar orang yang berkecimpung atau terkait dengan dunia pendidikan dan psikologi. Di sisi lain, seperti telah disebutkan di muka, kelemahan teori tes klasik, antara lain, adalah: (a) kemampuan peserta tes dinyatakan dalam variabel yang deskrit, dan (b) besarnya koefisien validitas dan koefisien reliabilitas suatu tes serta nilai parameter suatu butir soal tergantung kepada peserta yang dikenai suatu tes.

Karena munculnya teori respons butir dimaksudkan untuk menutup kelemahan-kelemahan yang ada pada teori tes klasik. maka keunggulan teori respons butir. antara lain. adalah: (a) lebih baik landasan teorinya dibandingkan dengan teori tes klasik. (b) kemampuan peserta tes dinyatakan dalam variabel yang kontinu. (c) tidak diperlukan tes paralel untuk menghitung koefisien reliabilitas (yang dalam teori respons butir disebut fungsi informasi), dan (d) besarnya koefisien reliabilitas suatu tes dan nilai parameter suatu butir soal tidak tergantung kepada peserta tes yang dikenai suatu tes. Namun demikian, penggunaan teori respons butir mengandung sejumlah kelemahan. antara lain. adalah: (a) memerlukan sampel besar untuk dapat menghasilkan parameter yang stabil. sehingga konsep teori respons butir tidak dapat diterapkan pada tingkat kelas. (b) diperlukan *software* (program komputer) yang andal untuk dapat melakukan estimasi parameter yang akurat. dan (c) belum diterima keberadaannya oleh sebagian besar orang yang berkecimpung di dunia pendidikan dan psikologi. terutama di Indonesia. Dewasa ini belum banyak *software* yang dapat digunakan oleh masyarakat untuk mengestimasi parameter butir soal berdasarkan teori respons butir. Andaikata sudah ada. belum banyak masyarakat yang dapat memakainya. Penolakan sebagian besar masyarakat Indonesia terhadap "konversi" nilai UAN yang dilakukan oleh pemerintah pada awal tahun duaribuan merupakan bukti bahwa keberadaan teori respons butir belum sepenuhnya dimengerti dan diterima oleh masyarakat Indonesia

BAHAN DISKUSI

1. Pada suatu ujian, skor yang diberikan kepada siswa dianggap sebagai skor tampak (X) atau skor sebenarnya (T)? Mengapa?

2. Pada suatu universitas, untuk memberikan nilai pada mahasiswa diberikan empat kali ujian. yaitu UKD1, UKD2, UKD3. dan UKD4. Nilai akhir mahasiswa adalah rerata dari UKD-UKD tersebut. Dari sisi teori pengukuran, mengapa tidak cukup dengan menggunakan satu UKD saja. tetapi harus menggunakan 4 UKD?

3. Dari sisi pengukuran, setujukah Anda kalau Ujian Nasional itu hanya diberikan satu kali saja dalam setahun? Alasan apa kira-kira yang Pemerintah Republik Indonesia berikan, sehingga Pemerintah Republik Indonesia hanya memberikan Ujian Nasional satu kali saja dalam setahun?

4. Pada teori tes klasik, terdapat apa yang disebut dengan *true score*. Jika kita mengukur kemampuan aljabar. apakah *true score* untuk seorang

murid dapat diamati? Jika tidak dapat diamati, bagaimana mendapatkan *true score* murid tersebut?

5. Pada suatu ujian, tiba-tiba terjadi kebakaran, sehingga para siswa diminta segera menyelesaikan pekerjannya. Walaupun demikian, hasil ujian para siswa tetap diskor seperti biasanya. Kesalahan pengukuran yang terjadi bersifat acak atau sistematik? Mengapa?

6. Pada suatu ujian, terjadi kecurangan demikian rupa sehingga setiap siswa mendapatkan bocoran kunci jawaban. Walaupun demikian, hasil ujian para siswa tetap diskor seperti biasanya. Kesalahan pengukuran yang terjadi bersifat acak atau sistematik? Mengapa?

7. Pada suatu kontes, misalnya, Indonesia Mencari Bakat, mengapa jurinya tidak hanya satu orang tetapi tiga orang? Setujukah Anda kalau jurinya tidak tiga orang, tetapi 11 orang?

8. Dengan menggunakan formula $\rho_{XX'} = 1 - \frac{\sigma_E^2}{\sigma_X^2}$, tunjukkan bahwa $0 \leq \rho_{XX'} \leq 1$.

9. Seorang peneliti mengestimasi koefisien reliabilitas tes yang dibuatnya dengan menggunakan rumus KR-20. Orang tersebut memperoleh koefisien reliabilitas tesnya adalah 2,4. Kalau peneliti itu Anda, apa yang Anda lakukan? Mengapa?

10. Seseorang mengestimasi koefisien reliabilitas tes yang dibuatnya dengan menggunakan rumus KR-20. Orang tersebut memperoleh koefisien reliabilitas tesnya adalah −0,4. Kalau seseorang itu Anda, apa yang Anda lakukan? Mengapa?

11. Pada suatu ujian pilihan ganda, biasanya seseorang memperoleh skor 1 apabila jawabannya terhadap suatu butir benar dan memperoleh skor 0 apabila jawabannya terhadap butir tersebut salah. Jika misalnya penskorannya diubah dengan memberikan skor 5 jika jawabannya benar dan memberikan skor 1 jika jawabannya salah, setujukah Anda? Mengapa?

12. Misalnya terdapat 10 butir soal pilihan ganda. Butir nomor 1, 2, dan seterusnya berturut-turut mempunyai tingkat kesulitan 0,5, 0,7, 0,4,

0,5, 0,5, 0,6, 0,3, 0,5, 0,7, dan 0,5. Cara penskoranya adalah sebagai berikut.

$$\text{Skor} = \sum_{i=1}^{10} k_i P_i$$

dengan $k_i = 1$ jika butir ke-i dijawab benar. $k_i = 0$ jika butir ke-i dijawab salah dan P_i adalah tingkat kesulitan butir ke-i.

a. Parti menjawab benar nomor 1, 5, 6, 7, dan 8. Berapakah skor Parti?
b. Wanti menjawab benar nomor 7, 8, 9, dan 10. Berapakah skor Wanti?
c. Dibandingkan dengan cara pensekoran ujian pilihan ganda yang biasa dilakukan orang, keunggulan dan kelemahan apa yang terjadi pada cara penskoran itu?
c. Setujukan Anda dengan cara penskoran seperti itu, jika penskorannya dilakukan secara manual? Mengapa?
d. Setujukan Anda dengan cara penskoran seperti itu, jika penskorannya dilakukan dengan komputer, misalnya dengan membuat program aplikasi tertentu? Mengapa?

13. Pertanyaannya seperti pad soal Nomor 12, tetapi rumus penskorannya adalah sebagai berikut.

$$\text{Skor} = \sum_{i=1}^{10} iP_i \quad \text{dengan } P_i \text{ adalah tingkat kesulitan butir ke-i.}$$

14. Sebuah tes mempunyai koefisien reliabilitas sebesar 0.6 dengan variansi skor tampak sebesar 25. Berapakah variansi skor sebenarnya dan variansi kesalahan skornya? Berapakah kesalahan baku pengukurannya?

15. Pada suatu pengukuran dengan suatu tes tertentu, variansi skor sebenarnya adalah 16 dan variansi skor kesalahannya adalah 4. Berapakah koefisien reliabilitasnya? Berapakah kesalahan baku pengukurannya?

16. Setujukah Anda kalau ada orang mengatakan bahwa rumus KR-20 adalah rumus untuk **menghitung** koefisien reliabilitas suatu tes? Mengapa?

17. Setujukah Anda kalau ada orang mengatakan bahwa rumus KR-20 adalah rumus untuk mengestimasi koefisien reliabilitas suatu tes, bukan untuk menghitung koefisien reliabilitas? Mengapa?

18. Suatu instrumen yang mempunyai panjang 25 butir mempunyai koefisien reliabilitas 0,60. Jika tes tersebut diperpanjang menjadi 30 butir, berapa koefisien reliabilitas tes yang baru?

19. Suatu instrumen yang mempunyai panjang 25 butir mempunyai koefisien reliabilitas 0,60. Berapa butir harus ditambahkan agar instrumen tersebut mempunyai koefisien reliablitas sebesar 0,70?

20. Suatu instrumen yang mempunyai panjang 25 butir mempunyai koefisien reliabilitas 0,90. Berapa butir harus ditambahkan agar instrumen tersebut mempunyai koefisien reliablitas sebesar 1,00?

*****

BAB III
TES DAN PERSYARATANNYA

PENDAHULUAN

Tes didefinisikan sebagai seperangkat pertanyaan atau tugas yang direncanakan untuk memperoleh informasi tentang trait atau atribut pendidikan atau atribut psikologik tertentu yang setiap butir pertanyaan atau tugas tersebut mempunyai jawaban atau ketentuan yang dianggap benar (Asmawi Zainul & Noehl Nasution, 1995: 3). Dengan demikian setiap tes menuntut keharusan adanya respons dari peserta tes yang dapat disimpulkan sebagai suatu trait yang dimiliki oleh peserta tes. Respons dari peserta tes tersebut harus dapat dikategorikan sebagai respons yang benar atau respons yang salah. Jika ada pertanyaan atau tugas yang harus dikerjakan oleh seseorang, tetapi tidak ada jawaban atau cara mengerjakan yang benar atau salah, maka pertanyaan atau tugas tersebut bukanlah suatu tes.

Di sisi lain, AERA, APA, dan NCME (1999: 3) mendefinisikan tes sebagai "*an evaluative device or procedure in which sample of an examinee's behaviour in a specified domain is obtained and subsequently evaluated and scored using standardized process*". Berdasarkan definisi ini, tes adalah alat atau prosedur evaluatif di mana sampel perilaku peserta tes dari domain tertentu diambil dan kemudian dinilai dan diskor menggunakan proses yang baku (standar). Dengan demikian, ketika seseorang memberikan tes mengenai kemampuan aljabar, maka sebenarnya seseorang tersebut hanya mengambil sampel perilaku (dalam hal ini adalah kemampuan mengerjakan aljabar) dari peserta tes. Namun demikian, seperti halnya pada statistika inferensial, hasil yang diperoleh pada sampel itu diberlakukan secara umum pada populasinya[1].

---

[1] Ini berarti jika seorang siswa SMP mendapat nilai 100 pada Ujian Nasional SMP, maka harus diartikan bahwa nilai 100 itu diberlakukan pada populasinya. Artinya, harus diartikan

Setelah tes selesai disusun, maka pengembang tes wajib mengujicobakannya terlebih dulu sebelum digunakan sebagai alat untuk melakukan penilaian. Tujuan uji coba adalah untuk melihat apakah tes yang disusun telah memenuhi persyaratan sebagai tes yang baik atau belum. Analisis untuk melihat apakah suatu tes telah memenuhi persyaratan sebagai tes yang baik atau belum disebut analisis tes (atau analisis instrumen). Tes yang baik harus valid dan reliabel. Kecuali melakukan analisis instrumen, pengembang tes juga melakukan analisis butir instrumen. Pada tes prestasi, misalnya, analisis butir soal meliputi analisis untuk melihat: (1) memadai atau tidaknya tingkat kesukaran, (2) memadai atau tidaknya daya pembeda, dan (3) berfungsi atau tidaknya pengecoh (pada tipe pilihan ganda).

VALIDITAS

Banyak definisi mengenai validitas tes. Nunnaly (1978: 86) dan Allen dan Yen (1979, 95) mengatakan bahwa suatu tes disebut valid jika tes tersebut mengukur apa yang seharusnya diukur. Ini adalah definisi validitas yang banyak digunakan orang. Pada definisi tersebut, istilah validitas dikaitkan dengan instrumen, yaitu tes.

Namun demikian, banyak ahli yang mendefinisikan validitas dalam kaitannya dengan skor tes, seperti yang dikatakan oleh Guilford (1954: 398) bahwa istilah validitas menunjuk kepada sejauh mana skor tes dapat memprediksi kriteria yang telah ditentukan. Senada dengan itu, Cronbach (1971) seperti yang ditulis oleh Crocker dan Algina (1986:217) mendefinisikan validasi sebagai suatu proses di mana pengembang tes atau pengguna tes mengumpulkan bukti-bukti untuk mendukung berbagai jenis inferensi yang dapat ditarik dari skor tes.

Definisi lebih komprehensif dari validitas dikemukakan oleh Messick (1989: 13) sebagai berikut: "*validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rational support the adequacy and appropriateness of inference and actions based on test scores or other modes of assessment*". Pada sisi lain, *Standards for Educational and Psychological Testing* AERA, APA, dan NCME (1999: 9) mendefinisikan validitas sebagai "*the degree to which evidence and theory support the interpretation of test score entailed by proposed uses of test*".

Berdasar itu dapat dikatakan bahwa validitas adalah penilaian evaluatif terintegrasi, yang dilakukan oleh penilai mengenai seberapa jauh bukti-bukti empirik dan rasional teoritis mendukung ketepatan inferensi dan tindakan berdasar skor tes atau asesmen yang lain.

---

bahwa siswa tersebut mendapat nilai 100 untuk kemampuannya menguasai matematika SMP, tidak sekedar mendapat nilai 100 pada Ujian Nasionalnya saja.

Jenis-jenis Validitas

Konsep teoritik validitas berkembang dari tahun ke tahun. Pada mulanya validitas berkenaan dengan prediksi dari kriteria spesifik, seperti yang dikatakan oleh Guilford pada tahun 1946 bahwa tes adalah valid untuk sesuatu yang berkorelasi dengannya. Kemudian, fokus dari validitas adalah interpretasi dari skor tes. Pergeseran dari prediksi ke eksplanasi sebagai fokus dari validitas ini, menyebabkan bahwa penggunaan, rele-vansi, dan pentingnya prediksi tidak dapat diukur ketika tidak ada skor yang dapat dipakai untuk melakukan prediksi.

Pada tahun 1954, APA (*American Psychological Association*), menyatakan ada empat jenis validitas, yaitu: validitas isi, validitas prediktif, validitas konkuren, dan validitas konstruks. Kemudian, pada tahun 1966, APA mereduksinya menjadi tiga jenis (Messick, 1989:18), yaitu validitas isi (*content validity*), validitas berdasar kriteria (*criterion-related validity*), dan validitas konstruk (*construct validity*). Penggolongan validitas ke beberapa jenis tersebut didasarkan kepada tujuan khusus dari instrumen yang dikenakannya. *Pada 1966 Standards*, dikatakan bahwa validitas isi bertujuan untuk menentukan apakah yang ditampakkan secara individual dapat pula ditampakkan pada keseluruhan (*universe*) situasi; validitas berdasar kriteria bertujuan untuk memprediksi keadaan masa depan individual atau keadaannya sekarang berdasar beberapa variabel yang berbeda dengan tes yang ditempuhnya; dan validitas konstruk bertujuan untuk menarik kesimpulan mengenai tingkatan kualitas seseorang berdasarkan kepada kinerjanya dalam tes.

Walaupun ada tiga jenis validitas di atas, Cronbach (1984), seperti yang dinyatakan oleh Messick (1989:19), menekankan bahwa penggolongan validitas ke dalam tiga tipe tersebut tidaklah saling pilah. Cronbach mengatakan bahwa "*the end goal of validation is explanation and understanding*", sehingga dia sampai kepada kesimpulan bahwa "*the profession is coming around to the view that all validation is construct validation*". Jadi, Cronbach mengatakan bahwa pada dasarnya validitas adalah validitas konstruks.

Walaupun terdapat berbagai jenis validitas, tetapi untuk tes prestasi belajar, validitas yang paling tepat adalah valditas isi. Untuk skala sikap yang mengukur mengenai kecemasan, misalnya, maka disamping validitas isi, seharusnya dilakukan juga validitas konstruks.

Validitas isi

Pada beberapa instrumen, validitas bergantung kepada ketepatan pemilihan sampel atas domain atau isi tertentu suatu *behaviour* (tingkah laku). Jika ini yang dipakai sebagai acuan, maka validitas yang dibicarakan

adalah validitas isi[2]. Dengan demikian, suatu instrumen disebut valid menurut validitas isi apabila isi instrumen tersebut telah merupakan sampel yang representatif dari keseluruhan isi hal yang akan diukur. Validitas isi sering disebut validitas ahli.

Dikatakan oleh Nunnally (1978, 92) bahwa ada dua standar utama untuk meyakinkan adanya validitas isi, yaitu: (1) koleksi butir-butir soal yang representatif terhadap semestanya, dan (2) metode penyusunan tes yang masuk akal (*sensible*). Dalam tes prestasi belajar, untuk meyakinkan bahwa butir-butir soal telah mewakili tujuan pembelajaran atau kompetensi dasar tertentu, diperlukan adanya *outline* rinci, atau *blue-print* (kisi-kisi) yang memuat pertanyaan atau permasalahan apa saja yang harus diujikan. Dalam kasus-kasus seperti ini, penilaian kualitas kisi-kisi merupakan bagian penting untuk menilai validitas isi.

Untuk tes hasil belajar, supaya tes mempunyai validitas isi, harus diperhatikan hal-hal berikut.

(1) Bahan ujian (tes) harus merupakan sampel yang representatif[3] untuk mengukur sampai seberapa jauh tujuan pembelajaran tercapai ditinjau dari materi yang diajarkan maupun dari sudut proses belajar.
(2) Titik berat bahan yang diujikan harus seimbang dengan titik berat bahan yang telah diajarkan.
(3) Tidak diperlukan pengetahuan lain yang tidak atau belum diajarkan untuk menjawab soal-soal ujian dengan benar.

Untuk mempertinggi validitas isi, disarankan agar pembuat soal melalui langkah-langkah berikut[4].

(1) Mengidentifikasikan bahan-bahan yang telah diberikan beserta tujuan pembelajarannya atau indikator-indikator dari kompetensi dasar yang diukur.
(2) Membuat kisi-kisi dari soal tes yang akan ditulis. Cara yang ditempuh adalah membuat tabel dua jalan yang memuat isi pokok bahasan (atau indikator) yang akan diukur dan aspek tingkah laku yang akan dinilai (menurut Taksonomi Bloom, misalnya)

---

[2] Untuk skripsi atau tesis, menurut penulis cukup dilakukan validasi isi saja.
[3] Misalnya diberikan Ujian Nasional Matematika tingkat SMP yang berupa tes pilihan ganda dengan 5 alternatif jawaban. Banyaknya butir soal adalah 40 butir dengan lama waktu pengerjaan 120 menit. Perhatikanlah bahwa 40 butir yang diujikan itu merupakan sampel dari populasi yang seharusnya diujikan. Populasi yang seharusnya diujikan (diukur) adalah kemampuan Matematika SMP yang dipelajari siswa selama mereka sekolah 3 tahun di SMP, yang kalau diujikan seluruhnya memerlukan waktu berhari-hari.
[4] Langkah ini dilakukan kalau uji coba untuk mencari parameter butir (daya beda dan tingkat kesulitan) dan koefisien reliabilitas tidak memungkinkan.

(3) Menyusun soal tes beserta kuncinya. Dalam hal ini menyusun kunci sesaat setelah menulis soal tes sangat dianjurkan.
(4) Menelaah soal tes sebelum dicetak. Penelaahan ini akan lebih baik apabila dilakukan oleh satu tim yang terdiri dari ahli-ahli yang relevan.

Jika misalnya peneliti membuat suatu tes hasil belajar untuk mengukur variabel terikatnya (yaitu prestasi belajar), maka validasi isi dilakukan langkah-langkah berikut.

(1) Mengidentifikasikan bahan-bahan yang telah diberikan beserta tujuan pembelajarannya atau indikator-indikator dari kompetensi dasar yang diukur. Pada penelitian, biasanya hanya satu atau kompetensi dasar tertentu, namun terdiri dari sejumlah besar (misalnya 15 buah) indikator.
(2) Merencanakan berapa butir yang seharusnya dipakai untuk mengukur variabel terikat. Untuk menghindari butir-butir yang harus dibuang setelah dilakukan uji coba, maka banyaknya butir yang diujicoba harus lebih banyak daripada banyaknya butir yang diperlukan, misalnya diberi lebihan 25%.
(3) Membuat kisi-kisi dari soal tes yang akan ditulis. Cara yang ditempuh adalah membuat tabel dua jalan yang memuat isi pokok bahasan yang akan diukur dan aspek tingkah laku yang akan dinilai (menurut Taksonomi Bloom, misalnya). Pada praktik penelitian, kisi-kisi ini adalah kisi-kisi untuk tes yang diujicobakan.
(4) Menyusun soal tes beserta kuncinya. Dalam hal ini menyusun kunci sesaat setelah menulis soal tes sangat dianjurkan.
(5) Menyerahkan soal tes, kunci jawaban, beserta cara penyelesaiannya kepada validator (*expert*) untuk dimintakan komentarnya. Perhatikanlah bahwa tugas validator bukan untuk menentukan butir mana yang harus dibuang, tetapi tugas validator adalah memberikan masukan kepada peneliti mengenai soal tes yang dibuatnya.

Kadang-kadang pengembang tes menyatakan bahwa penulisan butir-butir tes dengan baik dari domain-domain spesifik (di kisi-kisi) yang disusun secara cermat telah memenuhi validitas isi. Tetapi ini sebenarnya bukan merupakan kegiatan validasi isi. Kegiatan validasi isi adalah serangkaian kegiatan yang berlangsung setelah bentuk awal instrumen telah selesai ditulis. Kegiatan ini dapat dilakukan oleh pengembang tes ataupun oleh pengguna tes yang tidak terlibat dalam penyusunan tes.

Untuk menilai apakah suatu instrumen mempunyai validitas isi yang tinggi, yang dilakukan adalah melalui *experts judgment* (penilaian yang dilakukan oleh para pakar). Dalam hal ini para penilai (yang sering disebut *subject-mater experts*), melakukan dua hal pokok. Pertama, para penilai menilai apakah kisi-kisi yang dibuat oleh pengembang tes telah menunjuk-

kan bahwa klasifikasi kisi-kisi telah mewakili isi (substansi) yang akan diukur atau telah sesuai dengan konsep yang telah didefinisikan. Kedua, para penilai menilai apakah masing-masing butir tes yang telah disusun cocok atau relevan dengan klasifikasi kisi-kisi yang ditentukan[5]. Cara ini sering disebut *relevance ratings* (penilaian berdasar relevansi). Pada cara ini, biasanya, kepada para penilai diberikan suatu rentangan skala tertentu (misalnya 1-10, di mana 1 menunjukkan sangat-sangat tidak relevan dan 10 menunjukkan sangat-sangat relevan, atau hanya dua kemungkinan yaitu baik dan tidak baik), kemudian ditentukan suatu rating (yang merupakan rataan dari para penilai) untuk masing-masing klasifikasi kisi-kisi dan masing-masing butir soal. Hasil dari *relevance ratings* ini dapat berupa mo-difikasi kisi-kisi, atau modifikasi butir soal, atau keduanya oleh pengem-bang tes berdasar masukan dari validator.

Secara singkat, pada tingkat minimum, langkah-langkah dalam melakukan validasi isi, Crocker dan Algina menawarkan adanya empat langkah berikut.

(1) Mendefinisikan domain kinerja yang akan diukur (pada tes prestasi dapat berupa serangkaian tujuan pembelajaran atau pokok-pokok bahasan atau sejumlah kompetensi dasar yang diwujudkan dalam kisi-kisi),
(2) Membentuk sebuah panel yang ahli (*qualified*) dalam domain-domain tersebut,
(3) Menyediakan kerangka terstruktur untuk proses pencocokan butir-butir soal dengan domain performans yang terkait (kerangka terstruktur ini biasanya berwujud tabel-tabel atau matriks-matriks yang biasanya disebut Lembar Validasi), dan
(4) Mengumpulkan data dan menyimpulkan berdasar data yang diperoleh dari proses pencocokan pada Langkah (3).

Allen dan Yen (1979:95-96) membedakan validitas isi menjadi dua tipe, yaitu: (1) validitas tampang (*face validity*) dan (2) validitas logik (*logic validity*) atau validitas sampling (*sampling validity*). Validitas tampang dipenuhi apabila terdapat similaritas (kesesuaian) antara hasil tes dengan *trait* (kemampuan) yang relevan yang diukur dengan tes tersebut. Misalnya, suatu tes aritmetika mempunyai validitas tampang apabila tes tersebut mengukur kinerja peserta tes dalam melakukan pengerjaan aritmetika. Di sisi lain, validitas logik dipenuhi apabila *behaviour* yang diukur oleh tes dan disain logik dari butir-butir tes telah mencakup aspek-aspek

---

[5] Pada kegiatan validasi isi (atau validasi ahli) ada dua bagian. Bagian pertama adalah penelaahan kisi-kisi dan bagian kedua adalah penelaahan butir-butir soal kaitannya dengan kisi-kisi. Namun demikian, pada skripsi dan tesis mahasiswa, bagian yang pertama sering tidak dilakukan.

penting dalam domainnya. Validitas logik ini sangat esensial dalam pengembangan tes prestasi. Biasanya, yang disebut dengan validitas isi pada umumnya adalah validitas logik menurut Allen dan Yen.

Dalam perkembangannya, validitas isi menjadi kontroversial sebab banyak pakar pengembang tes yang mendefinisikan validitas dalam arti yang terkait dengan inferensi yang ditarik dari skor tes (lihat definisi dari Messick di muka). Pada hal studi pada validitas isi jarang yang bertumpu pada data skor tes. Biasanya, isi suatu tes divalidasi melalui metode subjektif seperti misalnya meminta penilai untuk memberi rating (skala) pada butir-butir soal apakah sesuai dengan klasifikasi kisi-kisi. Oleh karena itu, banyak yang mengusulkan penggantian nama validitas isi dengan nama lain yang lebih cocok, misalnya relevansi isi (*content relevance*), atau representasi isi (*content representation*), atau keterwakilan isi (*content representativeness*) (Sireci dan Geisinger, 1992:17).

Berikut ini adalah sebuah contoh lembar kerja untuk validasi isi.

Contoh 3.1.

Berikut ini adalah contoh Lembar Validasi untuk melihat kecocokan kisi-kisi dengan Kemampuan (kompetensi dasar) yang diukur

Petunjuk:

Perhatikan kisi-kisi yang telah dibuat oleh pengembang tes. Berikan komentar mengenai kisi-kisi tersebut dalam hubungannya dengan kompetensi dasar yang akan diukur, misalnya dalam kaitannya dengan hal-hal berikut.

a. Apakah kompetensi dasar yang akan diukur telah lengkap?
b. Jika terlalu banyak, kompetensi dasar saja yang harus dikurangi, dan jika terlalu sedikit, kompetensi dasar apa yang perlu ditambahkan.

Komentar Validator:

Contoh 3.2

Berikut ini adalah contoh tabel untuk melihat kesesuaian butir soal dengan kisi-kisi

Petunjuk:

Berilah tanda check pada kolom yang sesuai, jika butir soal telah memenuhi kriteria yang disebutkan. Jika tidak sesuai, berilah tanda silang, dan berikan komentar perbaikan mengenai butir soal tersebut pada lembar yang diberikan. Jika sekiranya lembar yang diberikan tidak mencukupi. Bapak/ Ibu validator dapat menambahkan pada lembar tersendiri.

Tabel 3.1. Contoh Lembar Validasi suatu Tes

No	Kriteria penelahaan[6]	Nomor Butir			
		1	2	...	40
Segi Materi (Substansi)					
1	Butir soal sesuai dengan kompetensi dasar atau indikator yang ingin dicapai				
2	Materi pada butir soal telah dipelajari oleh siswa				
3	Kunci jawaban pada butir soal telah benar				
Segi Konstruksi					
4	Pokok soal dirumuskan dengan singkat dan jelas				
5	Pokok soal bebas dari pernyataan yang dapat menimbulkan penafsiran ganda				
6	Jawaban butir soal ini tidak tergantung kepada jawaban butir soal yang lain				
7	Pengecoh butir soal sudah disusun dengan baik				
Segi Bahasa					
8	Butir soal menggunakan bahasa Indonesia yang baik dan benar				
9	Butir soal menggunakan bahasa yang komunikatif				
10	Butir soal tidak menggunakan bahasa atau istilah yang berlaku pada daerah tertentu				

Saran Masukan Validator:

Mohon dituliskan di bawah ini

1.

2.

3.

Validitas berdasar kriteria

Oleh Allen dan Yen (1979:97) dikatakan bahwa validitas berdasar kriteria (*criterion-related validity*) digunakan ketika skor tes dapat dihu-

---

[6] Kriteria penelaahan yang ditampilkan pada buku ini sekedar contoh. Para pembaca dapat mengembangkannya sendiri sesuai dengann tujuan penilaian.

bungkan dengan sebuah kriteria tertentu. Dalam hal ini kriteria adalah tingkah laku tertentu yang skor tes dapat digunakan untuk memprediksinya.

Dengan demikian, validitas berdasar kriteria adalah validitas yang ditinjau dari segi hubungan dengan alat pengukur lain yang dipandang sebagai kriteria untuk menentukan tinggi rendahnya validitas alat ukur yang sedang dipersoalkan.

Validitas semacam ini lebih menekankan pada kriterianya, bukan pada instrumennya itu sendiri. Berbeda dengan validitas isi, validitas berdasar kriteria ini menggunakan teknik-teknik empiris hubungan antara skor instrumen yang dipersoalkan dengan kriteria luar, sehingga identifikasi kriteria menjadi penting. Beberapa ciri yang harus dimiliki oleh suatu ukuran kriteria adalah relevansi, reliabel, dan bebas dari bias.

Ciri pertama adalah relevansi. Peneliti harus menilai apakah kriteria yang telah dipilih itu benar-benar menggambarkan ciri-ciri yang tepat dari tingkah laku yang diselidiki. Jika kriteria tersebut tidak mencerminkan atribut yang sedang diteliti, maka akan tidak ada artinya menggunakan kriteria tersebut. Ciri kedua adalah reliabilitas. Hal ini berarti bahwa kriteria tersebut harus merupakan ukuran yang ajeg bagi atribut tersebut, dari waktu ke waktu dan dari satu situasi ke situasi yang lain. Apabila kriteria itu sendiri tidak konsisten, maka tidak dapat diharapkan bahwa alat ukur yang dipersoalkan akan memberikan keajegan. Ciri ketiga adalah bebas dari bias. Ini berarti bahwa pemberian skor pada suatu kriteria hendaknya tidak dipengaruhi oleh faktor-faktor selain penampilan sebenarnya pada kriteria itu.

Misalnya, agar supaya sebuah tes penerimaan pegawai dapat mempunyai validitas berdasar kriteria, skor tes penerimaan tersebut harus berkorelasi dengan kriteria tertentu, misalnya efektivitas kerja (*job effectiveness*). Contoh lain, sebuah tes masuk perguruan tinggi mempunyai validitas berdasar kriteria apabila skor hasil tes berkorelasi dengan suatu kriteria tertentu, misalnya indeks prestasi mahasiswa.

Tinggi rendahnya indeks validitas berdasar kriteria biasanya dinyatakan oleh koefisien korelasi antara skor tes (prediktor) dengan skor kriteria. Salah satu koefisien korelasi yang dapat dipakai adalah koefisien korelasi momen produk dari Karl Pearson yang dirumuskan berikut.

$$r_{xy} = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}}$$

dengan r_{xy} adalah koefisien validitas, X adalah skor tes, dan Y adalah skor kriteria.

Contoh 3.3

Seorang peneliti ingin menghitung koefisien validitas tes masuk perguruan tinggi. Sebagai kriteria untuk menentukan validitas tes tersebut adalah IPK (Indeks Prestasi Kumulatif) setelah mahasiswa lulus. Misalnya datanya adalah sebagai berikut.

Tabel 3.2. Nilai Skor Tes Masuk dan IPK 10 Mahasiswa

No	Nama Mhsw	Skor Tes Masuk	IPK
1	Aa	45	3,45
2	Bb	65	3,65
3	Cc	85	4.00
4	Dd	65	3,54
5	Ee	75	3,63
6	Ff	60	3,65
7	Gg	55	3,50
8	Hh	45	3,24
9	Ii	75	3,60
10	Jj	95	4,00

Jawab:

Dengan memisalkan Skor Tes Masuk sebagai X dan IPK sebagai Y, dibuat tabel kerja berikut.

Tabel 3.3. Tabel Kerja untuk Mencari Koefisien Validitas Tes Masuk

No	X	Y	X^2	Y^2	XY
1	45	3,45	2025	11,9025	6986,25
2	65	3,65	4225	13,3225	15421,25
3	85	4,00	7225	16,0000	28900,00
4	65	3,54	4225	12,5316	14956,50
5	75	3,63	5625	13,1769	20418,75
6	60	3,65	3600	13,3225	13140,00
7	55	3,50	3025	12,2500	10587,50
8	45	3,24	2025	10,4976	6561,00
9	75	3,60	5625	12,9600	20250,00
10	95	4,00	9025	16,0000	36100,00
Jumlah	$\sum X =$ 665	$\sum Y =$ 36,26	$\sum X^2 =$ 46625	$\sum Y^2 =$ 131,9636	$\sum XY =$ 173321,25

$$r_{xy} = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}}$$

$$= \frac{(10)(17332,25) - (665)(36,26)}{\sqrt{((10)(46625 - 665^2)((10)(1319636) - 36.26^2)}}$$

$$= \frac{1709099,6}{\sqrt{(24025(4,848)}} = 0,501$$

Jadi, koefisien validitas tes masuk tersebut adalah 0.501. Koefisien validitas ini tergolong kecil. Dikatakan bahwa tes masuk tersebut tidak mempunyai validitas yang baik.

Secara umum, desain untuk melakukan validasi berdasar kriteria adalah sebagai berikut (Crocker dan Algina, 1986:224).

(1) Identifikasikan *behaviour* kriteria yang cocok dan cara untuk mengukur *behaviour* tersebut,
(2) Identifikasikan sampel dari peserta tes yang dapat mewakili peserta tes yang sesungguhnya akan dikenai tes,
(3) Selenggarakan tes dan simpanlah skor dari sampel peserta tes,
(4) Ketika kriteria sudah diperoleh, lakukan pengukuran kinerja pada kriteria tersebut untuk setiap sampel peserta, dan
(5) Tentukan koefisien korelasi antara skor tes dengan skor kinerja sebagai kriteria, yang koefisien korelasi tersebut merupakan koefisien validitas.

Validitas berdasar kriteria dikelompokkan menjadi dua jenis, yaitu validitas prediktif (*predictive validity*) dan validitas konkuren (*concurent validity*).

Pada validitas prediktif, skor kriteria yang dipakai untuk memprediksi *behaviour* tidak tersedia ketika tes dilakukan, tetapi tersedia di kemudian hari. Jadi, pada kasus ini, kriterianya tidak tersedia pada saat tes berlangsung, namun kriterianya baru dapat ditentukan setelah selang waktu tertentu. Misalnya, pada tes penerimaan karyawan baru, kriterianya adalah kinerja pegawai. Indeks kinerja pegawai ini baru dapat ditentukan setelah pegawai yang diterima diobservasi kualitas kinerjanya setelah beberapa waktu, misalnya setelah satu tahun bekerja. Pada kasus seperti ini, indeks validitas prediktif hanya dapat dilihat pada mereka yang telah diterima sebagai pegawai, yang indeks validitas berdasar cara ini pada umumnya berada di bawah indeks validitas yang sebenarnya. Tentu saja validitas prediktif ini menjadi mahal dan memerlukan waktu yang lama.

Di sisi lain, pada validitas konkuren, kriteria yang dipakai untuk mengkorelasikan hasil tes telah ada (atau dapat dicari) pada saat tes berlangsung. Misalnya, untuk melihat validitas suatu tes yang baru disusun,

digunakan tes standar yang telah diakui mempunyai indeks validitas yang tinggi. Kedua tes tersebut dikenakan pada sekelompok siswa yang sama (atau dua kelompok siswa yang sama kondisinya) pada saat yang sama (atau hampir bersamaan), kemudian dicari koefisien korelasi antara skor tes yang baru disusun dengan skor tes standar. Apabila koefisien korelasinya tinggi, maka dapat disimpulkan bahwa tes yang baru disusun mempunyai indeks validitas konkuren yang tinggi.

Validitas konstruks

Validitas konstruks (*construct validity*) adalah jenis validitas yang paling akhir dikembangkan orang (Cronbach dan Meehl, 1955, pada Allen dan Yen, 1979:108). Validitas konstruks suatu tes adalah suatu ukuran yang mengukur konstruk teoretis atau *trait* (kemampuan) yang seharusnya diukur (*the degree to which it measures the theoretical construct or trait that it was designed to measure*). Berdasarkan teori terbaru yang terkait dengan *trait* yang akan diukur, pengembang tes membuat prediksi tentang bagaimana skor tes berperilaku dalam berbagai situasi. Prediksi ini kemudian diuji. Jika prediksi ini tidak didukung oleh data, maka tes tersebut tidak mempunyai validitas konstruks seperti yang diteorikan.

Jika prediksi yang dibuat tidak didukung data, maka ada 3 kemungkinan yang terjadi, yaitu: (1) pelaksanaan eksperimen kurang baik, (2) teori pendukungnya kurang baik, dan (3) tes tidak mengukur *trait* yang diinginkan.

Dalam mengembangkan validitas konstruks, beberapa prediksi yang dapat diuji, antara lain sebagai (Allen & Yen, 1979: 108) berikut.

1. Perbedaan Kelompok. Jika teori yang mendukungnya mengatakan bahwa ada perbedaan antara kelompok yang dipikirkan, maka pengembang tes harus mengumpulkan data terkait dengan perbedaan tersebut dan mengujinya dengan statistik. Misalnya berdasarkan teori, harus ada perbedaan antara anak-anak dan remaja pada tes *social maturity*, maka pengembang tes menguji perbedaan skor tes pada kelompok anak-anak dan kelompok remaja.

2. Perubahan. Jika teori pendukungnya mengatakan bahwa ada perubahan skor seiring dengan berjalannya waktu atau perubahan yang lain, maka pengembang tes harus menguji perubahan tersebut. Misalnya tes yang mengukur *oral-communication skills* harus menghasilkan skor yang lebih tinggi seiring dengan pertumbuhan anak-anak.

[7] Untuk disertasi, kecuali disyaratkan pengujian validitas isi (ahli), menurut penulis buku ini, diperlukan pengujian dengan validitas konstruks.

3. Korelasi. Teori pendukungnya bisa jadi memunculkan adanya korelasi positif, negatif, atau nol antara berbagai variabel yang mungkin muncul. Kalau teori mengatakan seperti itu, maka pengembang tes harus mengujinya secara statistik. Misalnya tes yang mengukur *short-term memory* menurut teori harus berkorelasi positif dengan umur, tetapi tidak berkorelasi dengan jenis kelamin.

4. Proses. Misalnya terdapat tes *mathematical-reasoning* yang berisi soal cerita (*word problems*) yang menggunakan kata-kata atau kalimat-kalimat yang sangat sukar. Untuk melakukan validitas konstruk pada tes tersebut. pengembang tes harus menguji apakah teori yang melandasinya didukung oleh data. Misalnya membedakan antara peserta tes yang *vocabulary*nya baik dan yang tidak.

Validitas Faktorial

Validitas faktorial (*factorial validity*) adalah salah satu bentuk validitas konstruks yang dibangun melalui analisis faktor. Validitas jenis inilah yang sering dipakai untuk melakukan validitas konstruks. Analisis faktor adalah suatu istilah yang menyatakan sejumlah besar prosedur matematik untuk melakukan analisis mengenai interrelasi antara sejumlah variabel dan menjelaskan interrelasi tersebut dalam sejumlah variabel yang lebih sedikit, yang disebut faktor. Faktor adalah variabel hipotetik (ada yang menyebut variabel laten) yang mempengaruhi skor pada satu atau lebih variabel amatan.

Ada dua jenis analisis faktorial. Yang pertama adalah analisis faktor eksploratori (*exploratory factor analysis*), sedangkan yang kedua adalah analisis faktor konfirmatori (*confirmatory factor analysis*). Analisis yang dipakai untuk menguji validitas konstruks adalah analisis faktor konfirmatori, yang pada dasarnya membandingkan pembagian faktor-faktor ketika merencanakan tes dengan faktor-faktor yang diperoleh dengan analisis faktor konfirmatori. Jika keduanya menunjukkan hal yang sama, maka tes tersebut telah memenuhi validitas konstruks.

Untuk melakukan validasi konstruks dengan analisis faktor konfirmatori, dapat digunakan software tertentu, misalnya Lisrel atau AMOS.

RELIABILITAS

Suatu instrumen disebut reliabel apabila hasil pengukuran dengan instrumen tersebut adalah sama jika sekiranya pengukuran tersebut dilakukan pada orang yang sama pada waktu yang berlainan atau pada orang-orang yang berlainan (tetapi mempunyai kondisi yang sama) pada waktu yang sama atau pada waktu yang berlainan. Dengan kata lain, sebuah tes

disebut reliabel jika seseorang diuji dengan tes tersebut beberapa kali akan menghasilkan skor yang sama atau beberapa orang yang kemampuannya sama diuji dengan tes tersebut akan menghasilkan skor yang sama. Kata reliabel sering disebut dengan nama lain, misalnya terpercaya, terandalkan, ajeg, stabil, konsisten, dan lain sebagainya.

Reliabilitas menunjuk kepada konsistensi hasil pengukuran jika dilakukan pengukuran berulang-ulang pada individu-individu atau kelompok-kelompok dalam suatu populasi (*AERA, APA, & NCME*, 1999: 25). Ini berarti, keterandalan suatu tes menunjuk kepada besarnya kesalahan pengukuran yang dihasilkan oleh tes tersebut. Semakin besar koefisien keterandalan suatu tes akan semakin kecil kesalahan pengukurannya (Djemari Mardapi, dkk, 2002: 113).

Pada umumnya tidak pernah didapatkan instrumen yang mempunyai reliabilitas sempurna, sebab setiap kali mengadakan pengukuran dengan alat yang sama terhadap subjek yang sama biasanya diperoleh hasil yang berbeda. Hal ini disebabkan adanya kesalahan (yang mungkin juga ditimbulkan oleh instrumen itu sendiri atau ditimbulkan oleh orang yang menggunakan instrumen itu), yang akibatnya skor yang diperoleh dari suatu subjek bukanlah skor yang sebenarnya, melainkan skor yang sudah ditambah dengan kesalahannya. Dengan demikian, sebuah instrumen mempunyai reliabilitas yang tinggi apabila derajat kesalahannya kecil.

Mengacu kepada adanya kesalahan tersebut, biasanya orang mengatakan bahwa hasil pengukuran dapat dipercaya apabila dalam beberapa kali pelaksanaan pengukuran terhadap subjek yang sama atau kelompok subjek yang sama diperoleh hasil pengukuran yang relatif sama, selama aspek yang diukur dalam diri subjek atau kelompok subjek itu memang tidak berubah. Tentu saja suatu instrumen tidak harus dikenakan beberapa kali kepada subjek yang sama. Jika suatu instrumen tidak dikenakan kepada subjek (atau kelompok subjek) yang sama, suatu instrumen menghasilkan hasil pengukuran yang dapat dipercaya jika dikenakan kepada subjek (atau kelompok subjek) yang berlainan, tetapi dengan kondisi yang sama, menghasilkan hasil pengukuran yang relatif sama. Alat ukur tinggi badan, misalnya mempunyai reliabilitas yang tinggi, sebab jika dipakai untuk mengukur seseorang akan menghasilkan tinggi badan yang sama, sekalipun pengukuran itu dilakukan berulang-ulang. Jika alat ukur tinggi badan itu dipakai untuk mengukur orang yang berlainan, tetapi mempunyai tinggi badan yang sama, pasti akan menghasilkan tinggi badan yang sama.

Merujuk uraian pada Bab II, pada teori tes klasik, koefisien reliabilitas instrumen dinyatakan dengan $\rho_{XX'}$. Terdapat banyak rumus untuk koefisien reliabilitas tersebut, di antaranya adalah:

$$\rho_{XX'} = 1 - \frac{\sigma_e^2}{\sigma_X^2}$$

Karena $\sigma_X^2 = \sigma_T^2 - \sigma_e^2$, dapat dibuktikan bahwa rentang koefisien reliabilitas adalah:

$$0 \leq \rho_{XX'} \leq 1$$

Suatu instrumen disebut reliabel jika $\frac{\sigma_e}{\sigma_X} \leq \frac{1}{2}$, sehingga dari sini dapat dikatakan bahwa suatu instrumen disebut reliabel jika $\rho_{XX'} \geq \frac{3}{4}$ atau dengan mengambil penyederhanaan, suatu instrumen disebut reliabel jika $\rho_{XX'} \geq 0.70$.

→ koefisien reliabilitasnya.

Pada dasarnya koefisien reliabilitas tidak dapat dihitung, karena data mengenai error (e) tidak diketahui, sehingga variansinya pun tidak diketahui. Oleh karena itu, orang mengembangkan berbagai cara untuk **mengestimasi** koefisien reliabilitas.

Nilai reliabilitas instrumen yg paling tinggi adl

Kesalahan Baku Pengukuran

Perhatikan salah satu formula koefisien reliabilitas berikut ini.

$$\rho_{XX'} = 1 - \frac{\sigma_e^2}{\sigma_X^2}$$

Berdasarkan formula itu dapat diperoleh $\sigma_e = \sigma_X \sqrt{1 - \rho_{XX'}}$. Besaran σ_E disebut kesalahan baku pengukuran (*the standard error of measurement*). Jika koefisien reliabilitas diestimasi dari sampel, maka kesalahan pengukuran dirumuskan oleh $s_e = s_X \sqrt{1 - r_{XX'}}$ atau $s_e = s_X \sqrt{1 - r_{11}}$ jika koefisien reliabitas dinyatakan dengan r_{11}.

Contoh 3.4

Dengan menggunakan rumus KR-20, seorang peneliti memperoleh r_{11} = 0,85. Variansi skor yang diperoleh adalah 14. Berapakah kesalahan baku pengukurannya?

Jawab:

$r_{11} = 0{,}85$; $s_X = 14$; $s_e = ?$

$$s_e = s_X\sqrt{1-r_{11}} = 14\sqrt{1-0.95} = 3.13$$

Jadi kesalahan baku pengukurannya adalah 3,13.

METODE UNTUK MENGESTIMASI KOEFISIEN RELIABILITAS

Metode yang digunakan untuk mengestimasi koefisien reliabilitas instrumen (terutama tes hasil belajar) dapat dikelompokkan menjadi tiga golongan besar, yaitu: (a) metode satu kali tes, (b) metode tes ulang, dan (c) metode bentuk sejajar (paralel). Metode mana yang sebaiknya dipakai, tidak ada aturan baku. Namun, biasanya orang memilih metode satu kali tes, sebab metode ini mudah dilakukan dan berbiaya murah dibandingkan dengan dua pendekatan yang lainnya.

Perlu diketahui bahwa tiga macam metode tersebut menghasilkan koefisien reliabilitas yang berbeda-beda. Dianjurkan kepada pengembang tes untuk mencantumkan meode dan teknik mana yang dipakai. Pencantuman tersebut sangat penting untuk menghindari (mengurangi) salah tafsir dari pihak yang menggunakan tes tersebut.

Metode Satu Kali Tes

Metode ini disebut juga *single-test method* atau *single-trial method*. Dengan metode ini pengembang tes hanya melakukan pengukuran (menggunakan instrumen yang dipersoalkan reliabilitasnya) kepada sekelompok subjek satu kali saja. Reliabilitas yang didasarkan atas metode ini biasanya disebut *internal consistency reliability*.

Metode ini merupakan metode yang paling banyak dipakai karena merupakan metode yang paling ekonomis dan paling praktis. Beberapa teknik yang sering digunakan dalam metode satu kali tes adalah sebagai berikut.

Teknik Spearman-Brown (pilgan)

Teknik ini dikenal pula dengan teknik belah-dua, sebab dalam menentukan koefisien reliabilitasnya, soal tes dikelompokkan menjadi dua bagian yang sebanding (paralel, setara). Cara yang banyak digunakan ialah membelah alat pengukur menjadi butir-butir yang bernomor genap menjadi satu bagian dan butir-butir yang bernomor gasal menjadi bagian yang lain. Oleh karena itu, teknik ini sering disebut teknik gasal-genap (*odd-even technique*).

Kadang-kadang pembagiannya mengacu kepada nomor urut butirnya. Misalnya suatu soal tes terdiri dari 40 butir soal, maka butir-butir soal nomor 1 sampai dengan 20 menjadi bagian pertama, sedangkan butir-butir

nomor 21 sampai dengan 40 menjadi bagian kedua. Bagian pertama dan kedua merupakan bagian yang saling paralel.

Instrumen (yang sebenarnya terdiri dari dua bagian) itu diberikan kepada sekelompok subjek. Dengan sendirinya masing-masing subjek akan mempunyai dua buah skor, yaitu skor bagian pertama dan skor bagian kedua. Koefisien korelasi antara dua macam skor itu disebut $r_{\frac{1}{2}\frac{1}{2}}$.

Spearman dan Brown merumuskan koefisien reliabilitas instrumen sebagai berikut.

$$r_{11} = \frac{2r_{\frac{1}{2}\frac{1}{2}}}{1+r_{\frac{1}{2}\frac{1}{2}}} \qquad 3.1$$

dengan r_{11} adalah koefisien reliabilitas instrumen dan $r_{\frac{1}{2}\frac{1}{2}}$ adalah koefisien korelasi antara skor bagian pertama dan bagian kedua.

Perhatikanlah bahwa rumus pada Persamaan 3.1 merupakan rumus Sperman-Brown yang ditulis pada Bab II dengan mengambil N = 2.

Untuk menggunakan rumus Spearman-Brown, ada beberapa syarat yang harus dipenuhi, yaitu:

(a) Dua belahan yang diciptakan harus merupakan dua tes yang paralel.

(b) Banyaknya butir instrumen harus genap.

(c) Instrumen yang dicari reliabilitasnya harus homogen.

Teknik Flanagan

Kelemahan dari teknik Spearman-Brown ialah bahwa syarat pertama tersebut di atas sulit dipenuhi. Untuk menutupi kelemahan itu, Flanagan menciptakan rumus sebagai berikut.

$$r_{11} = 2\left(1 - \frac{s_1^2 + s_2^2}{s_t^2}\right) \qquad 3.2$$

dengan r_{11} adalah koefisien reliabilitas instrumen, s_1^2 adalah variansi instrumen belahan pertama, s_2^2 adalah variansi instrumen belahan kedua, dan s_t^2 adalah variansi instrumen total.

Jika dikenakan kepada populasi, rumus tersebut berubah menjadi berikut.

$$\rho_{11} = 2(1 - \frac{\sigma_1^2 + \sigma_2^2}{\sigma_t^2})$$

Teknik Rulon

Teknik lain yang mendasarkan pada pembelahan alat pengukur menjadi dua bagian yang sama ialah teknik yang dikembangkan oleh Rulon. Teknik ini berpangkal kepada dasar pemikiran bahwa perbedaan antara skor subjek uji coba pada bagian pertama dan skor subjek uji coba pada bagian kedua adalah karena kesalahan pengukuran. Oleh karena itu, variansi yang diperoleh berdasarkan perbedaan tersebut dapat dipandang sebagai variansi skor kesalahan pada model X = T + e.

Rumus yang dikemukakan oleh Rulon adalah sebagai berikut.

$$r_{11} = 1 - \frac{s_d^2}{s_t^2} \qquad (3.3)$$

dengan r_{11} adalah koefisien reliabilitas instrumen, s_d^2 adalah variansi perbedaan skor antara dua belahan, dan s_t^2 adalah variansi skor total.

Perhatikanlah bahwa rumus pada Persamaan 3.3 merupakan turunan dari rumus koefisien reliabilitas $\rho_{XX'} = 1 - \frac{\sigma_e^2}{\sigma_X^2}$ pada Bab II.

Jika dikenakan kepada populasi, rumus Rulon tersebut berubah menjadi berikut.

$$\rho_{11} = 1 - \frac{\sigma_d^2}{\sigma_t^2}$$

Teknik Kuder-Richardson

Kuder dan Richardson merasa tidak puas dengan teknik belah dua. Mereka menganggap bahwa pembelahan instrumen menjadi dua bagian bukan merupakan teknik yang baik untuk mencari koefisien reliabilitas. Hal ini disebabkan dalam praktik, pembelahan menjadi dua bagian dapat dilakukan dengan bermacam-macam cara yang biasanya memperoleh hasil yang berbeda.

Untuk menghindari hal ini, Kuder dan Richardson tidak membelah menjadi dua, melainkan memperhatikan banyaknya butir dan memperhatikan banyaknya subjek yang menjawab benar pada tiap-tiap butir. Ini

berarti bahwa teknik Kuder-Richardson mendasarkan kepada analisis masing-masing butir.

Namun perlu diingat bahwa teknik ini hanya dapat dipakai untuk instrumen yang dikhotomus (setiap butir hanya mempunyai dua kategori skor yaitu 1 atau 0, seperti pada misalnya tes pilihan berganda, yang diskor 1 jika benar dan diskor 0 jika salah). Untuk instrumen skala sikap dengan skala Likert, teknik ini tidak dapat dipakai.

Rumus dari Kuder-Richardson berbentuk sebagai berikut.

$$r_{11} = \left(\frac{n}{n-1}\right)\left(\frac{s_t^2 - \Sigma p_i q_i}{s_t^2}\right)$$

dengan r_{11} adalah koefisien reliabilitas instrumen, n adalah banyaknya butir instrumen, p_i adalah proporsi banyaknya subjek yang menjawab benar pada butir ke-i, $q_i = 1 - p_i$, dan s_t^2 adalah variansi untuk skor total

Contoh 3.5

Misalnya terdapat 10 butir soal yang diujicobakan kepada 8 siswa dengan data sebagai berikut.

Tabel 3.4. Sebaran Skor untuk 8 Mahasiswa pada 10 Butir Soal

No	Nama	Nomor Butir Soal									
		1	2	3	4	5	6	7	8	9	10
1	Aa	1	0	1	0	1	1	1	1	1	1
2	Bb	1	0	1	0	1	1	0	1	1	1
3	Cc	1	0	1	0	1	1	0	1	1	1
4	Dd	1	0	1	0	0	1	1	1	0	1
5	Ee	1	0	0	1	0	1	1	0	1	0
6	Ff	1	0	0	1	1	0	0	1	0	0
7	Gg	1	0	0	1	0	0	1	0	0	0
8	Hh	1	0	0	1	0	0	0	0	0	1

Keterangan: jika skor pada butir tertentu adalah 1 berarti peserta tes yang bersangkutan menjawab benar butir tersebut dan jika skor butir tersebut 0 berarti peserta tes tersebut menjawab salah pada butir tersebut. Misalnya peserta tes yang bernama Aa benar menjawab keseluruhan butir, kecuali butir nomor 2 dan 4.

Estimasi koefisien reliabilitas tes tersebut dengan KR-20.

Jawab:
Untuk masing-masing peserta tes dihitung skor totalnya, kemudian dibuat tabel kerja seperti pada tabel berikut ini.

Tabel 3.5. Tabel Kerja untuk Mengestimasi Koefisien Reliabilitas

No	Nama	Nomor Butir Soal										Skor Total
		1	2	3	4	5	6	7	8	9	10	
1	Aa	1	0	1	0	1	1	1	1	1	1	8
2	Bb	1	0	1	0	1	1	0	1	1	1	7
3	Cc	1	0	1	0	1	1	0	1	1	1	7
4	Dd	1	0	1	0	0	1	1	1	0	1	6
5	Ee	1	0	0	1	0	1	1	0	1	0	5
6	Ff	1	0	0	1	1	0	0	1	0	0	4
7	Gg	1	0	0	1	0	0	1	0	0	0	3
8	Hh	1	0	0	1	0	0	0	0	0	1	3
	p	1	0	0,5	0,5	0,5	0,63	0,5	0,63	0,5	0,63	
	q	0	1	0,5	0,5	0,5	0,38	0,5	0,38	0,5	0,38	
	pq	0	0	0,25	0,25	0,25	0,23	0,3	0,23	0,25	0,23	$\Sigma pq =$ 1,95

Setelah dihitung, diperoleh $s_t^2 = 3,69$, sehingga:

$$r_{11} = \left(\frac{n}{n-1}\right)\left(\frac{s_t^2 - \Sigma p_i q_i}{s_t^2}\right) = \left(\frac{10}{9}\right)\left(\frac{3,69 - 1,95}{3,69}\right) = 0,523$$

Berdasarkan perhitungan tersebut, diperoleh koefisien reliabilitas tes sebesar 0,523.

Pada rumus di atas, jika datanya dianggap merupakan data populasi, maka rumus dari Kuder-Richardson berbentuk sebagai berikut.

$$\rho_{11} = \left(\frac{n}{n-1}\right)\left(\frac{\sigma_t^2 - \Sigma p_i q_i}{\sigma_t^2}\right)$$

dengan ρ_{11} adalah koefisien reliabilitas instrumen, n adalah banyaknya butir instrumen, p_i adalah proporsi banyaknya subjek yang menjawab benar pada butir ke-i, $q_i = 1 - p_i$, dan σ_t^2 adalah variansi skor total.

Di samping rumus KR-20. Kuder dan Richardson juga mengemukakan rumusnya yang lain, yang disebut rumus KR-21, sebagai berikut.

$$r_{11} = \left(\frac{n}{n-1}\right)\left(1 - \frac{R(n-R)}{ns_t^2}\right)$$

dengan r_{11} adalah koefisien reliabilitas instrumen, n adalah banyaknya butir instrumen, R adalah rerata skor total, dan s_t^2 adalah variansi skor total

Hasil penghitungan dengan KR-20 dan dengan KR-21 akan menghasilkan koefisien reliabilitas yang kurang lebih sama besarnya.

Teknik Alpha

Teknik alpha (koefisien alpha) ini dikembangkan pertama kali oleh Cronbach pada tahun 1951, dan karenanya sering disebut teknik Cronbach alpha. Berbeda dengan teknik Kuder-Richarson, teknik alpha dapat dipakai untuk instrumen yang tidak dikotomus (misalnya pada angket atau tes uraian).

Pada teknik ini, sebuah tes dapat dibelah menjadi beberapa bagian, misalnya k bagian (dengan $k \leq n$, n adalah banyaknya butir soal). Pada praktiknya, instrumen dapat dibelah menjadi n bagian, yang berarti masing-masing bagian terdiri dari satu butir saja. Pada teknik ini, masing-masing bagian dicari variansi skornya. Juga dicari variansi skor totalnya. Kemudian, koefisien reliabilitas dihitung dengan rumus berikut.

$$r_{11} = \left(\frac{n}{n-1}\right)\left(1 - \frac{\Sigma s_i^2}{s_t^2}\right) \qquad 3.4$$

dengan r_{11} adalah koefisien reliabilitas instrumen, n adalah banyaknya butir instrumen, s_i^2 adalah variansi belahan ke-i, i = 1, 2, ..., k ($k \leq n$) atau variansi butir ke-i, i = 1, 2, 3, 4,...., n , dan s_t^2 adalah variansi skor total yang diperoleh subjek uji coba. Rumus 3.4 tersebut sering disebut rumus Cronbach Alpha.

Pada rumus di atas, jika datanya dianggap merupakan data populasi, maka rumus dari Cronbach alpha berbentuk sebagai berikut.

$$\rho_{11} = \left(\frac{n}{n-1}\right)\left(1 - \frac{\Sigma \sigma_i^2}{\sigma_t^2}\right)$$

Contoh 3.6

Misalnya terdapat 10 butir soal yang diujicobakan kepada 8 siswa dengan data sebagai berikut.

Tabel 3.6. Sebaran Skor dari 8 Peserta Tes pada 10 Butir Soal

No	Nama	Nomor Butir Soal									
		1	2	3	4	5	6	7	8	9	10
1	Aa	1	0	1	0	1	1	1	1	1	1
2	Bb	1	0	1	0	1	1	0	1	1	1
3	Cc	1	0	1	0	1	1	0	1	1	1
4	Dd	1	0	1	0	0	1	1	1	0	1
5	Ee	1	0	0	1	0	1	1	0	1	0
6	Ff	1	0	0	1	1	0	0	1	0	0
7	Gg	1	0	0	1	0	0	1	0	0	0
8	Hh	1	0	0	1	0	0	0	0	0	1

Estimasi koefisien reliabilitas tes tersebut dengan menggunakan rumus Cronbach-Alpha.

Jawab:
Dicari skor totalnya, lalu dibuat tabel kerja sebagai berikut.

Tabel 3.7. Tabel Kerja untuk Mengestimasi Koefisien Reliabilitas

No	Nama	Nomor Butir Soal										Skor Total
		1	2	3	4	5	6	7	8	9	10	
1	Aa	1	0	1	0	1	1	1	1	1	1	8
2	Bb	1	0	1	0	1	1	0	1	1	1	7
3	Cc	1	0	1	0	1	1	0	1	1	1	7
4	Dd	1	0	1	0	0	1	1	1	0	1	6
5	Ee	1	0	0	1	0	1	1	0	1	0	5
6	Ff	1	0	0	1	1	0	0	1	0	0	4
7	Gg	1	0	0	1	0	0	1	0	0	0	3
8	Hh	1	0	0	1	0	0	0	0	0	1	3
	s_i^2	0	0	0,29	0,29	0,29	0,27	0,29	0,27	0,29	0,27	s_t^2 = 3,69

Setelah dihitung, diperoleh $\sum s_i^2 = 2,26$ dan $s_t^2 = 3,69$, sehingga diperoleh:

$$r_{11} = \left(\frac{n}{n-1}\right)\left(1 - \frac{\Sigma s_i^2}{s_t^2}\right) = \left(\frac{10}{9}\right)\left(1 - \frac{2.26}{3,69}\right) = 0,431$$

Jadi, koefisien reliabilitas tesnya adalah 9.431.

Contoh 3.8

Misalnya terdapat 5 butir soal bentuk uraian yang diujicobakan kepada 8 siswa dengan data sebagai berikut.

Tabel 3.8. Sebaran Skor dari 8 Peserta Tes pada 5 Butir Soal Uraian

No	Nama Siswa	Nomor Butir Soal				
		1	2	3	4	5
1	Aa	9	8	9	7	5
2	Bb	8	8	8	7	6
3	Cc	7	8	8	6	7
4	Dd	6	6	7	8	8
5	Ee	5	6	8	4	7
6	Ff	4	5	8	3	6
7	Gg	3	5	7	2	5
8	Hh	3	4	8	2	6

Jawab:

Lebih dulu dihitung skor totalnya, variansi masing-masing butir, dan variansi skor totalnya, sehingga diperoleh tabel kerja seperti pada Tabel 3.9.

Berdasarkan bilangan-bilangan pada Tabel 3.9, diperoleh koefisien reliabitas berdasar rumus Cronbach-Alpha sebagai berikut.

$$r_{11} = \left(\frac{n}{n-1}\right)\left(1 - \frac{\Sigma s_i^2}{s_t^2}\right)$$

$$= \left(\frac{8}{7}\right)\left(1 - \frac{14,946}{42,411}\right) = 0,740$$

Berarti koefisien reliabilitas tes tersebut adalah 0,740.

Tabel 3.9. Tabel Kerja untuk Menghitung Koefisien Reliabilitas

No	Nama	Skor Butir					Skor Total
		1	2	3	4	5	
1	Aa	9	8	9	7	5	38
2	Bb	8	8	8	7	6	37
3	Cc	7	8	8	6	7	36
4	Dd	6	6	7	8	8	35
5	Ee	5	6	8	4	7	30
6	Ff	4	5	8	3	6	26
7	Gg	3	5	7	2	5	22
8	Hh	3	4	8	2	6	23
Variansi Butir		5,13	2,50	0,411	5,839	1,071	$s_t^2 =$ 42,411
Jumlah Variansi Butir		$\sum_{i=1}^{5} s_i^2 = 14,946$					

Metode Tes Ulang

Metode ini disebut juga *test-re-test method*. Pada metode ini dilakukan pengukuran kepada sekelompok subjek dua kali dengan alat pengukur yang sama dalam waktu yang hampir bersamaan. Koefisien reliabilitasnya dihitung dengan mencari koefisien korelasi antara hasil pengukuran yang pertama dengan yang kedua. Rumus yang dipakai biasanya adalah rumus korelasi momen produk dari Karl Pearson.

Asumsi yang dipakai pada metode ini ialah tidak ada penambahan dan/atau pengurangan kemampuan peserta tes dan pelaksanaan dua kali tes tersebut dalam kondisi psikologis yang sama. Misalnya pelaksanaan tes pada hari pertama dalam suasana yang menyenangkan dan pelaksanaan tes pada hari kedua dalam suasana yang menegangkan, maka situasi seperti ini tidak menguntungkan untuk mengestimasi koefisien reliabilitas tes.

Misalnya ingin dicari koefisien reliabilitas dari tes A. Maka tes A tersebut diberikan kepada sekelompok siswa dua kali, misalnya hari ini dan besok pagi. Misalnya X adalah skor tes A pada hari ini dan Y adalah skor tes A pada besok pagi. Maka koefisien reliabilitasnya dicari dengan rumus berikut:

$$r_{11} = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}}$$

dengan r_{11} adalah koefisien reliabiltas. X adalah skor tes A pada hari pertama dan Y adalah skor tes A kedua.

Contoh 3.7

Misalnya sebuah tes diberikan kepada 10 siswa dua kali, yaitu pada 15 April 2014 dan 16 April 2014. Skor (total) mereka tampak pada Tabel 3.10. Lakukan estimasi koefisien reliabilitas tes tersebut dengan menggunakan metode tes ulang.

Tabel 3.10. Skor 10 Siswa pada Ujian 15 dan 16 April 20

No	Nama Siswa	Skor Tgl 15-04-24	Skor Tgl 16-04-12
1	Kk	68	70
2	Ll	73	72
3	Mm	45	46
4	Nn	90	92
5	Oo	86	85
6	Pp	75	78
7	Qq	84	80
8	Rr	85	95
9	Ss	34	36
10	Tt	46	47

Jawab:

Untuk mengestimasi koefisien reliabilitasnya diasumsikan skor tanggal 15 April 2014 sebagai X dan skor pada 16 April 2014 sebagai Y. Dibuat tabel kerja sebagai berikut.

Tabel 3.11. Tabel Kerja untuk Mencari Koefisien Reliabilitas

No	X	Y	X^2	Y^2	XY
1	68	70	4624	4900	4760
2	73	72	5329	5184	5256
3	45	46	2025	2116	2070
4	90	92	8100	8464	8280
5	86	85	7396	7225	7310
6	75	78	5625	6084	5850
7	84	80	7056	6400	6720
8	85	95	7225	9025	8075
9	34	36	1156	1296	1224
10	46	47	2116	2209	2162
Jumlah	$\sum X =$ 686	$\sum Y =$ 701	$\sum X^2 =$ 50652	$\sum Y^2 =$ 52903	$\sum XY =$ 51707

$$r_{11} = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}}$$

$$= \frac{(10)(51707) - (686)(701)}{\sqrt{\{(10)(50652) - 686^2\}\{(10)(52903) - 701^2\}}}$$

$$= \frac{36184}{\sqrt{(35924)(37629)}} = 0.98$$

Jadi koefisien reliabikitas tes tersebut adalah 0.98.

Metode Bentuk Paralel (Sejajar)

Metode ini disebut juga *parallel-form method*, *equivalent method*, atau *alternate forms*. Pada metode ini, dibuat dua buah instrumen yang paralel. Untuk menentukan reliabilitasnya, maka kedua instrumen tersebut diberikan kepada sekelompok subjek secara berturut-turut. Kemudian, hasil pengukuran dari instrumen tersebut dicari koefisien korelasinya. Koefisien korelasi tersebut sekaligus menentukan koefisien reliabilitas instrumen. Rumus yang biasanya digunakan adalah rumus korelasi momen produk dari Karl Pearson.

Misalnya ingin dicari koefisien reliabilitas dari tes A. Maka dibuat tes B (yang berbeda dengan tes A) yang paralel dengan tes A. Dua tes diberikan kepada sekelompok siswa secara berurutan. Misalnya X adalah skor tes A dan Y adalah skor tes B. Maka koefisien reliabilitasnya dicari dengan rumus berikut:

$$r_{11} = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}}$$

dengan r_{11} adalah koefisien reliabiltas, X adalah skor tes A, dan Y adalah skor tes B.

Metode ini sebenarnya berlandaskan pada pendefinisian koefisien reliabilitas yang dibicarakan di Bab II. Namun demikian, metode ini jarang dipakai, karena orang harus membuat dua tes paralel yang pada praktiknya tidak mudah dilakukan.

Reliabilitas Antarpenilai (Inter-Rater Reliability)

Jika skoring suatu penilaian sangat bersandar kepada subjektivitas penilai (misanya pada tes kinerja atau tes pada ranah psikomotor, atau pada tes bentuk uraian), maka reliabilitas antar-penilai perlu dipertimbangkan untuk digunakan. Ada dua cara untuk mencari koefisien reliabilitas berdasarkan reliabilitas antarpenilai tersebut. Cara pertama adalah dengan men-

cari koefisien korelasi antar nilai yang diberikan oleh dua penilai. Cara kedua adalah dengan melihat berapa persen kesesuaian yang diberikan oleh dua penilai (dalam arti dua penilai keduanya memberikan nilai atau skor yang sama).

Contoh 3.8

Misalnya terdapat 10 siswa yang diminta untuk membuat puisi. Dua orang penilai yang saling independen (penilai X dan penilai Y) diminta menilai puisi kesepuluh siswa tersebut berdasarkan rubrik yang diberikan, mulai dari 1 sampai dengan 5, dengan 1 = sangat jelek dan 5 = sangat baik. Nilai yang diberikan adalah sebagai berikut.

Tabel 3.12. Nilai Dua Penilai terhadap 10 Siswa dalam Membuat Puisi

Nama Penilai	Nilai anak ke-									
	1	2	3	4	5	6	7	8	9	10
Penilai X	4	5	3	4	5	4	3	2	3	4
Penilai Y	3	4	3	5	4	5	2	3	5	4

Cara Pertama:

Dicari koefisien korelasi antara nilai dari penilai X dan nilai dari penilai Y sebagai berikut. Diperoleh:

$$\sum X = 37; \sum Y = 38; \sum X^2 = 145; \sum Y^2 = 154; \text{dan} \sum XY = 144,$$

sehingga diperoleh:

$$r_{11} = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}}$$

$$= \frac{(10)(144) - (37)(38)}{\sqrt{\{(10)(145) - 37^2\}\{(10)(154) - 38^2\}}}$$

$$= \frac{34}{\sqrt{(81)(96)}} = 0{,}385$$

Jadi koefisien reliabilitasnya adalah 0,385.

Cara Kedua:

Banyaknya penilaian yang cocok antara penilai 1 dan penilai 2 ada 2 buah (yaitu ketika menilai siswa ke-3 dan ke-10). Banyaknya siswa yang dinilai seluruhnya ada 10 siswa. Jadi, koefisien reliabilitas antar-penilai adalah $r_{11} = \frac{2}{10} = 0{,}200$

Kadang-kadang dimodifikasi dengan memperkenankan adanya selisih 1, artinya penilaian masih dianggap cocok atau sesuai, walaupun selisih antara nilai yang diberikan penilai sama dengan 1. Misalnya, penilaian pada siswa ke-1 dianggap sesuai, walaupun penilai X memberi nilai 4 dan penilai Y memberi nilai 3. Hanya penilaian siswa ke-9 sajalah yang dianggap tidak cocok (penilai X memberi nilai 3 dan penilai Y memberi nilai 5), sehingga koefisien reliabilitas antarpenilainya adalah: $r_{11} = \frac{9}{10} = 0,900$.

PENAFSIRAN KOEFISIEN RELIABILITAS

Beberapa rumus koefisien reliabilitas instrumen dikembangkan dari rumus koefisien korelasi momen produk dari Karl Pearson, misalnya pada metode bentuk tes ulang maupun tes paralel.

Setelah koefisien reliabilitas instrumen diperoleh, tidak tergantung kepada metode mana yang dipilih, lalu diadakan penafsiran terhadap koefisien reliabilitas tersebut untuk menentukan reliabel atau tidaknya tes yang dipersoalkan[8].

Seperti telah diuraikan pada Bab II, pada umumnya, suatu instrumen dikatakan reliabel apabila koefisien reliabilitasnya 0,70 atau lebih ($r_{11} \geq 0,70$).. Ini berarti, hasil pengukuran yang mempunyai koefisien reliabilitas sebesar 0,70 atau lebih cukup baik nilai kemanfaatannya, dalam arti instrumennya dapat dipakai untuk melakukan pengukuran.

Pada beberapa buku, untuk melihat apakah suatu instrumen reliabel atau tidak, dilakukan hal-hal berikut. Setelah diperoleh koefisien reliabilitas, kemudian dilakukan uji signifikansi pada tingkat signifikansi tertentu (misalnya 5%) dengan melakukan uji statistik terhadap koefisien reliabilitas yang diperoleh dengan uji t atau dengan membandingkan tabel r. Cara ini tidak tepat dengan beberapa alasan. Pertama, menentukan reliabel atau tidaknya suatu instrumen pada dasarnya bukan uji signifikansi. Kedua, signifikan tidaknya suatu uji sangat tergantung kepada nilai n (banyaknya peserta tes). Jika n sangat besar, nilai r yang kecil pun akan signifikan.

Jadi, untuk menentukan apakah instrumen itu reliabel atau tidak maka yang dilihat adalah besarnya nilai koefisien reliabiltasnya dilihat dari segi kemanfaatannya, bukan signifikan atau tidaknya koefisien reliabilitas tersebut. Sekalipun koefisien reliabilitas tersebut signifikan, tetapi kalau nilai-

---

[8] Penafsiran kepada koefisien reliabilitas dilakukan berbeda menurut beberapa ahli dan beberapa buku. Namun demikian, menurut penulis, penafsiran dengan melakukan uji signifikansi terhadap koefisien reliabilitas itu tidaklah tepat.

nya kecil, maka koefisien reliabilitas tersebut tidak bermakna untuk menggambarkan apakah instrumennya reliabel atau tidak.

FAKTOR-FAKTOR YANG MEMPENGARUHI RELIABILITAS

Ada beberapa faktor yang mempengaruhi koefisien reliabilitas dari instrumen yang berupa tes, di antaranya sebagai berikut.

(a) **Panjang Tes**. Pada umumnya semakin panjang tes (dalam arti cacah butirnya semakin banyak) semakin tinggi koefisien reliabilitasnya. Hal ini disebabkan tes yang cacah butirnya banyak akan memuat cukup banyak tingkah laku yang diukur. Secara matematis, kebenaran pendapat ini dapat dibuktikan dengan menggunakan rumus Spearman-Brown yang telah ditulis di Bab II.

Dalam konteks ini, maka tes bentuk uraian (yang biasanya cacah butirnya sedikit) cenderung mempunyai koefisien reliablitas yang rendah. Di sisi lain, tes bentuk pilihan ganda (yang biasanya cacah butirnya banyak, sekitar 40 butir) cenderung mempunyai koefisien reliabilitas yang tinggi.

(b) **Penyebaran Skor**. Koefisien reliabilitas dipengaruhi oleh penyebaran skor. Makin lebar penyebaran skor (dalam arti makin besar variansinya), makin tinggi koefisien reliabilitasnya. Hal ini disebabkan koefisien reliabilitas akan semakin tinggi apabila individu-individu cenderung tetap pada kedudukan relatifnya terhadap kelompoknya.

Dalam kaitannya dengan tingkat kesulitan, makin mendekat nilai tingkat kesulitan butir ke bilangan 0,5 (yang berarti penyebaran skor totalnya semakin besar), maka koefisien reliablitas tesnya semakin tinggi.

(c) **Tingkat Kesulitan Tes**. Seperti yang telah disebutkan pada Bagian (b), tes yang terlalu sulit atau terlalu mudah cenderung menurunkan koefisien reliabilitas. Hal ini disebabkan tes yang terlalu sulit atau terlalu mudah menghasilkan sebaran yang terbatas dan terkumpul di ujung bawah atau di ujung atas. Dengan alasan ini pula, maka butir soal yang baik dari sisi tingkat kesulitan, adalah butir soal yang tingkat kesulitannya berada di sekitar setengah.

(d) **Objektivitas**. Objektivitas suatu tes menunjukkan seberapa jauh dua orang yang mempunyai kemampuan yang sama mendapatkan skor yang sama pula. Dalam hal ini, skor yang diperoleh oleh subjek yang dikenai tes tidak dipengaruhi oleh keputusan dan perasaan orang yang menentukan skor. Tes yang objektivitasnya tinggi cenderung mempunyai koefisien reliabilitas yang tinggi pula.

Pada konteks ini, dari sisi penyekoran, tes bentuk uraian cenderung tidak objektif, sedangkan tes bentuk pilihan ganda cenderung objektif. Oleh karena itu, tes bentuk uraian cenderung mempunyai koefisien reliabilitas rendah, dan tes bentuk pilihan ganda cenderung mempunyai koefisien reliabilitas tinggi.

BAHAN DISKUSI

1. Misalnya Anda membuat tes dalam kaitannya dengan pembuatan skripsi atau tesis (berarti Anda adalah peneliti dalam hal ini). Variabel terikatnya adalah prestasi belajar Pokok Bahasan Persamaan dan Pertidaksamaan Kuadrat. Ada 25 butir tes yang Anda validasikan ke pakar (berarti Anda melakukan validitas isi). Setelah divalidasi oleh pakar, ternyata ada 5 butir yang harus digugurkan karena tidak memenuhi kriteria penelaahan tertentu. Benarkah tindakan validator menggugurkan lima butir Anda pada kegiatan validasi ahli? Mengapa?

2. Seorang peneliti ingin menghitung koefisien validitas tes masuk perguruan tinggi. Sebagai kriteria untuk menentukan validitas tes tersebut adalah lama studi mahasiswa (dalam satuan tahun). Datanya adalah sebagai berikut.

Tabel 3.13. Skor Tes Masuk dan Lama Kelulusan 10 Mahasiswa

No	Nama Mhsw	Skor Tes Masuk	Lama (dalam tahun)
1	Aa	45	2,5
2	Bb	65	3,5
3	Cc	85	4,5
4	Dd	65	3.0
5	Ee	75	4.0
6	Ff	60	3.0
7	Gg	55	2,5
8	Hh	45	2,0
9	Ii	75	3.5
10	Jj	95	4.5

a. Hitunglah koefisien validitasnya. 0,95

b. Berdasarkan perhitungan pada (a), validkah tes masuk tersebut? Mengapa? dibandingkan dgn R tabel df = n-1 α = 0,05

etemu R tabel 0,521

arena 0,95 > 0,521 maka data tsb valid.

3. Berikut ini adalah sebaran 10 peserta tes yang mengikuti tes dengan 20 butir tes (nomor 1 sampai dengan 20).

Tabel 3.14: Skor Sepuluh Mahasiswa unuk 20 Butir Soal

	Nomor Butir Soal																			
No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	0	1	1	0	0	1	0	1	0	1	1	1	1	0	1	1	0	1	1
2	1	1	1	0	1	0	1	0	1	0	1	0	1	1	0	1	1	0	1	0
3	0	1	1	1	0	0	1	0	1	0	1	1	0	1	0	1	1	0	1	1
4	1	1	1	0	1	0	1	0	1	0	1	1	1	0	0	1	1	0	1	0
5	0	0	1	1	0	0	1	0	1	0	1	0	1	1	0	1	1	0	1	1
6	1	0	1	0	1	0	1	0	1	0	1	1	0	1	0	1	1	0	1	0
7	0	1	1	1	0	0	1	0	1	0	1	0	1	1	0	1	1	0	1	1
8	0	1	1	0	1	0	1	0	1	0	1	1	1	1	0	1	1	0	1	0
9	1	1	1	1	0	0	1	0	1	0	1	1	0	1	0	1	1	0	1	1
10	1	1	1	0	1	0	1	0	1	0	1	0	1	1	0	1	1	0	1	0

a. Hitunglah koefisien reliabilitas tes tersebut dengan menggunakan teknik belah dua dengan menganggap butir-butir soal nomor 1 sampai dengan 10 sebagai belahan pertama dan butir-butir soal nomor 11 sampai dengan 20 sebagai belahan kedua.

b. Hitunglah koefisien reliabilitas tes tersebut dengan menggunakan teknik belah dua dengan menganggap butir-butir soal bernomor gasal sebagai belahan pertama dan butir-butir soal bernomor genap sebagai belahan kedua.

4. Diketahui data seperti pada Soal Nomor 3.

   a. Hitunglah koefisien reliabilitas tes tersebut dengan menggunakan rumus Cronbach-Alpha dengan menganggap butir-butir soal nomor 1 sampai dengan 10 sebagai belahan pertama dan butir-butir soal nomor 11 sampai dengan 20 sebagai belahan kedua.

   b. Hitunglah koefisien reliabilitas tes tersebut dengan menggunakan rumus Cronbach-Alpha dengan menganggap tes tersebut terdiri dari 20 belahan (yang berarti masing-masing belahan hanya terdiri dari satu butir soal).

5. Misalnya terdapat 5 butir soal bentuk uraian yang diujicobakan kepada 8 siswa dengan data seperti pada Tabel 3.15.

Tabel 3.15. Skor 8 Mahasiswa unuk 5 Butir Soal Uraian

No	Nama Siswa	Nomor Butir Soal				
		1	2	3	4	5
1	Aa	9	8	9	7	5
2	Bb	8	8	8	7	6
3	Cc	7	8	8	6	7
4	Dd	6	6	7	8	8
5	Ee	5	6	8	4	7
6	Ff	4	5	8	3	6
7	Gg	3	5	7	2	5
8	Hh	3	4	8	2	6

a. Jika dapat, hitunglah koefisien reliabilitas tersebut dengan rumus Croncbach-Alpa!
b. Jika dapat, hitunglah koefisien reliabilitas tersebut dengan rumus KR-20.
c. Hitunglah kesalahan baku pengukuranya.

6. Berikut ini adalah sebaran jawaban sepuluh mahasiswa dalam menjawab 20 butir soal pilihan ganda dengan lima pilihan jawaban,

Tabel 3.16. Skor 10 Mahasiswa unuk 20 Butir Soal Pilihan Ganda

No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	A	A	B	F	B	B	C	E	E	D	C	B	D	C	A	A	B	A	C	D
2	C	A	B	D	B	B	C	E	E	B	C	E	D	C	A	A	B	A	B	D
3	C	A	B	E	B	B	C	E	E	B	C	E	D	C	A	A	C	A	D	B
4	C	B	A	E	B	C	D	E	E	A	B	C	D	A	A	A	B	A	C	D
5	D	A	B	D	B	E	D	E	C	B	D	C	D	C	A	A	C	B	E	E
6	C	A	C	D	B	B	B	E	E	B	C	E	D	C	A	A	B	A	D	C
7	C	A	B	D	B	B	B	E	E	F	C	E	D	C	B	C	C	B	A	A
8	B	D	C	D	C	C	B	C	D	A	D	B	B	D	C	C	A	C	D	A
9	A	B	C	B	E	D	A	A	C	D	A	B	A	A	V	A	C	C	D	A
10	A	B	C	B	A	D	E	C	D	B	A	B	D	C	E	A	E	B	C	D
KJ	**C**	**C**	**B**	**D**	**B**	**D**	**E**	**E**	**E**	**B**	**C**	**A**	**D**	**C**	**A**	**A**	**B**	**A**	**C**	**E**

Keterangan: KJ = kunci jawaban

a. Carilah koefisien reliabilitas tes tersebut dengan rumus KR-20!
b. Hitunglah kesalahan baku pengukuran jika menggunakan rumus KR-20!
c. Dapatkah koefisien reliabilitas tes tersebut dicari dengan rumur Cronbach-Alpha? Mengapa?
d. Dapatkah koefisien reliabilitas tes tersebut dicari degan teknik belah dua? Mengapa?

Tabel 3.17. Kisi-kisi untuk Tes Pilihan Ganda

No	Pokok Bahasan	Indikator	Taksonomi Bloom dan Nomor Soal		
			C2	C3	C4. C5. C6
1	Persamaan Kuadrat	1.1. Menyelesaikan persamaan kuadrat dengan $a = 1$ dan $b = 0$	1. 2		
		1.2. Menyelesaikan persamaan kuadrat dengan $a = 1$ dan $b = 0$	3. 4		
		1.3. Menyelesaikan persamaan kuadrat dengan $a = 1$ dan $c = 0$	5. 6		
		1.4. Menyelesaikan persamaan kuadrat dengan $a = 1$ dan $c \neq 0$	7. 8		
		1.5. Menyelesaikan persamaan kuadrat dengan $a > 1$ dan $b = 0$	17. 18		
		1.6. Menyelesaikan persamaan kuadrat dengan $a > 1$ dan $b \neq 0$	19, 20		
		1.7. Menyelesaikan persamaan kuadrat dengan $a < 0$	21, 22		
		1.8. Menyelesaikan soal cerita daiam bentuk persamaan kuadrat		29	30,31, 35
2	2. Pertidaksama-an Kuadrat	2.1. Menyelesaikan pertidakasamaan kuadrat dengan $a = 1$ dan $b = 0$	9, 10		
		2.2. Menyelesaikan pertidak-samaan kuadrat dengan $a = 1$ dan $b \neq 0$	11, 12		
		2.3. Menyelesaikan perstidaksamaan kuadrat dengan $a = 1$ dan $c = 0$	13, 14		
		2.4. Menyelesaikan persamaan kuadrat dengan $a = 1$ dan $c \neq 0$	15. 16		
		2.5. Menyelesaikan pertidaksamaan kuadrat dengan $a > 1$ dan $b = 0$	23. 24		
		2.6. Menyelesaikan pertidaksamaan kuadrat dengan $a > 1$ dan $b \neq 0$	25. 26		
		2.7. Menyelesaikan pertidaksamaan kuadrat dengan $a < 0$	27. 28		
		1.8. Menyelesaikan scal cerita dalam bentuk pertidaksamaan kuadrat		32	33,34

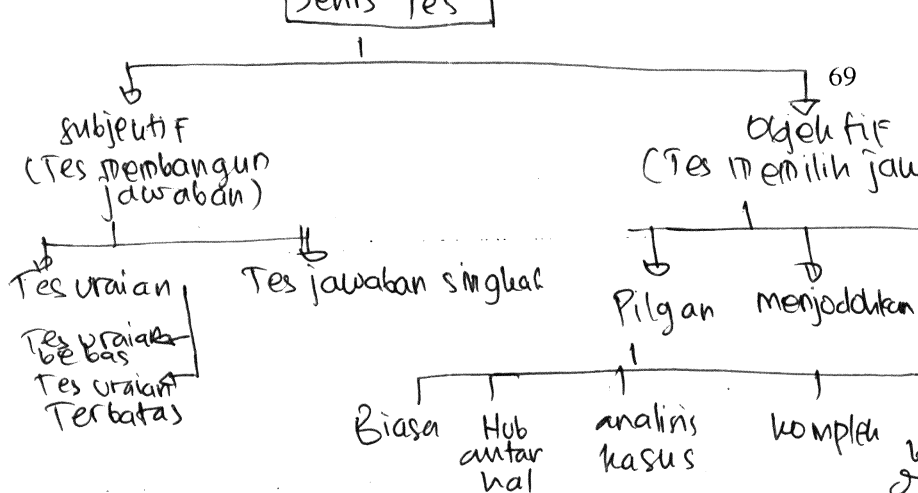
7. Misalnya Anda adalah seorang mahasiswa yang sedang menulis skripsi atau tesis. Anda menguji cobakan tes pilihan ganda pada Pokok Bahasan Persamaan Kuadrat dan Pertidaksamaan Kuadrat. Anda memerlukan 25 butir untuk mencari data penelitian nantinya, sedangkan pada saat uji coba dilakukan uji coba dengan 35 butir soal. Misalnya kisi-kisinya tampak seperti pada Tabel 3.17. Sebenarnya peneliti memerlukan 25 butir soal. Mengapa uji cobanya dengan 35 butir soal?

8. a. Misalnya Anda memerlukan 25 butir soal untuk mengukur suatu variabel terikat tertentu. Berapa butirkah yang Anda perlukan untuk uji coba?
   b. Misalnya Anda memerlukan 25 butir soal untuk mengukur vaiabel terikat tertentu, lalu Anda mengujicobakan hanya dengan 25 butir soal juga. Apa kelebihan dan kelemahan cara ini?

9. Perhatikan kisi-kisi pada Tabel 3.17. Buatlah lembar validasi untuk memvalidasi kisi-kisi dan butir-butir tes yang terkait dengan kisi-kisi pada Soal Nomor 7.

10. Misalnya Anda memerlukan 25 butir soal untuk mengukur variabel terikat tertentu. Kemudian Anda mengujicobakan 35 butir soal untuk menghindari kekurangan butir soal, karena ada kemungkinan ada butir-butir soal yang gugur (dibuang) ketika dianalisis setelah uji coba.
    a. Jika ternyata banyaknya butir soal yang baik adalah 30 butir, Anda menggunakan 25 butir atau 30 butir? Mengapa?
    b. Jika ternyata banyaknya butir soal yang baik adalah 20 butir, Anda menggunakan 20 butir atau 25 btutir? Mengapa?
    c. Jika ternyata banyaknya butir soal yang baik adalah 27 butir soal, yang Anda hitung koefisien reliabilitasnya yang 25 butir soal, 27 butir soal, atau 30 butir soal yang diujicobakan? Mengapa?

11. Perhatikan kembali kisi-kisi pada Tabel 3.12. Misalnya setiap indikator hanya Anda sediakan satu butir soal. Apa kelebihan dan kelemahan cara ini?

12. Ada program komputer untuk melakukan analisis instrumen dan analisis butir soal, baik untuk soal piligan ganda maupun untuk angket dengan skala Likert (dengan alternatif jawaban SS, S, N, TS, dan STS). Program itu namanya ITEMAN. Pada program itu, untuk mencari koefisien reliablitas instrumen yang dianalisis, digunakan rumus Cronbach-Alpha (coba Anda lakukan analisis dengan ITEMAN). Mengapa yang digunakan oleh ITEMAN adalah rumus Croncbah-Alpha, bukan KR-20?

*****

BAB IV
PENILAIAN RANAH KOGNITIF

PENDAHULUAN

Menurut Bloom, hasil belajar dapat dibedakan ke dalam tiga ranah (aspek, domain) yaitu hasil belajar ranah kognitif, ranah afektif, dan ranah psikomotor.

Pada bab ini dibicarakan berbagai jenis tes untuk ranah kognitif.

JENIS TES

Menurut bentuk pertanyaannya, pada umumnya orang membedakan tes ke dalam dua kelompok, yaitu tes membangun-jawaban (*constructed-response*) dan tes memilih-jawaban (*selected-response*). Tes membangun-jawaban sering disebut dengan tes subjektif, sedangkan tes memilih-jawaban sering disebut dengan tes objektif[1].

TES MEMBANGUN JAWABAN (*CONSTRUCTED-RESPONSE TEST*)

Termasuk ke dalam tes membangun jawaban adalah tes uraian (*essay test*) dan tes jawaban singkat (*short-answer test*)

TES URAIAN

Dulu kala, ujian-ujian sering dilaksanakan secara lisan dan ujiannya disebut ujian lisan. Pada ujian lisan, baik soal maupun jawabannya disampaikan secara lisan. Sampai dengan tahun seribu-sembilan-ratus-

---

[1] Literatur sekarang tidak menyebutnya sebagai tes objektif, sebab nama itu memberi arti bahwa tes di luar tes objektif tidaklah objektif.

tujuh-puluhan, di perguruan tinggi masih terdapat mata-mata kuliah yang diujikan secara lisan. Sampai sekarangpun, ujian skripsi, tesis, dan disertasi masih dilaksanakan dengan ujian lisan (*oral examination*).

Pada ujian lisan, kebanyakan pertanyaan yang diberikan penguji adalah tipe *constructed-response* di mana siswa atau mahasiswa diminta untuk memberikan penjelasan mengenai sesuatu yang ditanyakan, kadang disertai argumentasi yang mendukung jawaban itu.

Seiring dengan banyaknya siswa dan/atau mahasiswa yang terus bertambah dan kemajuan teknologi (misalnya dalam penyediaan kertas dan percetakan) ujian lisan tidak lagi populer, sebab *time-cosuming* dan berbiaya mahal. Dari sisi teori pengukuran dan pengujian, ujian lisan tidak baik, sebab tidak menyediakan kondisi yang seragam bagi semua peserta tes. Cara pemberian skornya juga merupakan permasalahan tersendiri. Oleh karena itu, untuk ujian matakuliah, sekarang ini sudah jarang dilakukan ujian lisan.

Salah satu tes membangun-jawaban (constructed respons) adalah tes uraian (*essay test*)[2]. Pada tipe ini, peserta tes diharapkan merumuskan jawaban sendiri dengan kata-kata sendiri. Jawaban tipe tes uraian dapat berupa jawaban pendek atau jawaban panjang, tergantung dari arah dan cakupan yang dikehendaki oleh butir tes. Jenis tes ini biasanya memuat permasalahan yang menuntut peserta tes untuk mengorganisir dan merumuskan jawabannya dengan menggunakan kata-kata, ide, dan/atau pemikirannya sendiri berdasar latar belakang pengetahuan yang dimilikinya.

Hal yang perlu diperhatikan dalam penyusunan soal tipe ini adalah bahwa rumusan permasalahannya hendaknya cukup jelas sehingga setiap peserta tes dapat menangkap permasalahannya dengan tepat seperti apa yang dimaksudkan oleh pembuat soal.

Untuk memperoleh tes uraian yang baik, perlu diperhatikan hal-hal berikut (Reynolds, Livingstode, dan Willson, 2010: 229-230).

1. Tulislah butir soalnya dalam kalimat yang jelas dan langsung (*clear and straightforward*). Termasuk dalam hal ini adalah seberapa panjang jawaban yang diharapkan (misalnya dalam 200 kata) dan berapa skor yang diberikan kepada butir soal itu (misalnya skornya 15 dari 100).
2. Perhatikan benar-benar lama waktu yang diberikan kepada peserta tes untuk menjawab seluruh butir soal yang diberikan. Kebanyakan guru *under-estimated* kepada lama waktu yang diberikan. Guru mengira siswanya dapat menyelesaikan seluruh butir soal dalam waktu 90 menit,

[2] Dulu orang sering menyebut tes uraian sebagai tes subjektif, karena hasil penyekorannya banyak dipengaruhi oleh pemeriksa. Namun demikian, sekarang ini, istilah tes subjektif dipandang tidak tepat, karena bisa memberikan konotasi bahwa hasil penyekorannya ditentukan oleh pemeriksanya, bukan oleh kualitas jawaban peserta tes.

misalnya, pada hal siswa yang pandai pun belum tentu dapat menyelesaikan seluruh butir soal dalam waktu 90 menit. Kalau perlu, pembuat soal mencoba mengerjakan seluruh butir soal yang ada, lalu menambahkan 25%nya untuk siswanya, karena siswa memerlukan waktu untuk membaca butir soal dan memerlukan waktu untuk mengorganisir jawaban.

3. Jangan memperbolehkan siswa memilih butir soal yang dikehendaki. Beberapa penguji memberikan perintah soal seperti: "Kerjakan 4 soal dari 5 soal yang disediakan". Perintah soal seperti ini seharusnya dihindari, karena siswa akan mengerjakan tes yang berlainan, dan karenanya tidak dapat diperbandingkan antara kemampuan siswa yang satu dengan yang lainnya. Kecuali itu, suruhan yang seperti ini mengganggu validitas isi.
4. Lebih baik diberikan butir soal pendek-pendek yang banyak daripada diberikan butir soal yang panjang tetapi sedikit. Hal ini disebabkan lebih banyak butir soal lebih reliabel dan lebih mewakili domainnya.
5. Jangan menulis butir soal yang lebih tepat disajikan dalam bentuk lain (misalnya bentuk pilihan ganda). Kalau misalnya soalnya hanya untuk aspek ingatan, lebih baik disajikan dalam bentuk pilihan ganda dibandingkan bentuk uraian.

Keunggulan tes uraian, antara lain: (1) menghendaki pengorganisasian jawaban, sehingga pada tes uraian dapat dilihat jalan pikiran peserta tes, (2) jawaban disampaikan berdasarkan kata-kata dan tulisannya sendiri, sehingga dapat dilihat kejernihan jalan pikiran peserta tes, (3) mudah menyusun soalnya, dan (4) dapat membedakan secara jelas kemampuan masing-masing siswa.

Di sisi lain, kelemahan tes uraian, antara lain: (1) bahan yang diliput terbatas, (2) waktu yang dipakai untuk menjawab soal tes uraian lama, (3) penilaian yang cenderung subjektif (cenderung dipengaruhi oleh penilai), dan (4) sukar dalam memberikan skor.

Untuk mengurangi kelemahan penggunaan tes uraian, dalam memeriksa tes uraian hendaknya diperhatikan hal-hal berikut.

1. Tetapkanlah dengan tepat hal-hal atau faktor-faktor yang diukur. Kemudian, penguji hendaknya hanya mengukur hal-hal atau faktor-faktor yang ditetapkannya tadi.
2. Bacalah dulu beberapa contoh jawaban untuk mendapatkan gambaran umum mengenai kualitas seluruh peserta tes.
3. Berdasarkan analisis pada langkah kedua, buatlah rubrik (kriteria pemberian skor) yang terkait dengan soal tersebut. Dalam membuat rubrik tersebut, penguji dianjurkan untuk membaca kembali catatan-catatan atau buku-buku yang dipakai sebagai referensi pembelajaran.

Termasuk dalam hal ini adalah menetapkan pokok-pokok penting yang harus ada untuk mendapatkan skor.

4. Periksalah setiap butir soal dalam satu waktu tertentu. artinya periksalah nomor butir soal yang sama untuk setiap siswa dalam satu waktu yang sama sebelum pemeriksaan nomor butir soal berikutnya.
5. Sedapat mungkin periksalah jawaban-jawaban soal tanpa mengetahui siapa penjawabnya.
6. Reliabilitas penilaian yang lebih besar diperoleh dengan jalan merata-ratakan skor yang diberikan oleh beberapa pemeriksa yang bekerja secara independen.

Tipe Tes Uraian

Tes uraian dapat dibedakan menjadi dua yaitu tes uraian bebas (*extended-response*) dan tes uraian terbatas (*resricted-response*).

Pada tes uraian bebas. peserta tes dapat dengan bebas menyatakan pendapat dan/atau penalarannya masing-masing. Boleh jadi, masing-masing peserta tes mengemukakan jawaban yang berbeda. walaupun mungkin sama-sama benarnya. Di sisi lain, pada tes uraian terbatas, jawaban siswa yang benar sudah dapat ditebak sebelumnya dengan variasi jawaban yang tidak banyak. Tes uraian untuk mata pelajaran Matematika, biasanya, termasuk ke dalam tes uraian terbatas.

Contoh 4.1

Berikut ini adalah contoh tes uraian bebas.

1. **Jelaskan, bagaimanakah pendapat Anda mengenai kualitas pembelajaran matematika di sekolah dasar sekarang ini?**
2. **Perlukah keterampilan menggunakan komputer diberikan kepada siswa-siswa sekolah dasar? Mengapa? Jelaskan pendapat Anda,**
3. **Manakah yang lebih tepat dilakukan pada siswa-siswa SMP, untuk mencari titik puncak suatu parabol. dengan menggambar grafiknya lebih dulu atau dengan menggunakan rumus? Jelaskan pendapat Anda.**

Pada tes uraian terbatas. walaupun jawaban dari peserta tes diurai menurut jalan pikiran masing-masing peserta tes. tetapi jawaban yang benar telah dapat diduga terlebih dulu. Jawaban yang benar dari masing-masing peserta tes relatif tidak berbeda, lebih-lebih untuk bidang eksakta.

Contoh 4.2

Berikut ini adalah tes uraian terbatas pada Matapelajaran Matematika.

1. **Dengan menggambar grafik fungsi kuadratnya terlebih dulu, selesaikan pertidaksamaan** $x^2 - 5x + 6 > 0$!

2. Diketahui $A = \begin{pmatrix} 4 & -9 \\ 3 & -4p \end{pmatrix}$, $B = \begin{pmatrix} 5p & -5 \\ 1 & 3 \end{pmatrix}$, dan $C = \begin{pmatrix} -10 & 8 \\ -4 & 6p \end{pmatrix}$. Jika $A - B = C^{-1}$, carilah p.

Cara Penskoran Tes Uraian

Tes uraian cenderung mempunyai reliabilitas dan validitas (isi) yang rendah. Untuk menutupi kekurangan itu, maka perlu dibuat cara penskoran (rubrik) yang jelas dan rinci. Keberadaan rubrik itu sekaligus juga untuk mengurangi subjektivitas penilai.

Pada dasarnya ada dua jenis rubrik, yaitu rubrik holistik (*holistic scoring rubric*) dan rubrik analitik (*analytic scoring rubric*). Rubrik holistik menilai berdasarkan kesan keseluruhan terhadap pekerjaan, sedangkan rubrik analitik berdasarkan penilaian terhadap bagian-bagiannya. Berikut ini adalah contoh dari Reynolds, Livingstode, dan Willson (2010: 235).

Contoh 4.3.

Diberikan butir soal berikut.

Bandingkan model inteligensi dari Thursone dan Gardner. Berilah contoh dimana perbedaan dan persamaannya.

Tabel 4.1. Contoh Rubrik Holistik

Klasifikasi	Deskripsi	Skor
bagus sekali	Siswa dapat menjelaskan dengan sangat bagus kedua metode secara akurat, dapat secara akurat menjelaskan perbedaan dan persamaannya, serta memberi cukup banyak contoh	5
bagus	Siswa dapat menjelaskan dengan bagus kedua model dan dapat menjelaskan perbedaan dan persamaannya.	4
di atas rata-rata	Siswa dapat menjelaskan kedua model dengan cukup baik tetapi ada beberapa kekurangan di sana-sini.	3
di bawah rata-rata	Siswa dapat menjelaskan kedua model, namun tidak dengan jelas. Beberapa informasi yang disampaikan tidak akurat.	2
Jelek	Siswa dapat menjelaskan secara terbatas kedua model, tetapi tidak dapat menyebutkan perbedaan dan persamaannya.	1
jelek sekali	Siswa tidak dapat menjelaskan sama sekali keberadaan kedua model.	0

Tabel 4.2. Contoh Rubrik Analitik

Bagian	Jelek (0)	Rata-rata (1)	Bagus (2)	Bagus Sekali (3)
Siswa dapat menjelaskan model Thurstone dengan baik	.....	.....	.....	.....
Siswa dapat menjelaskan model Gardner dengan baik	.....	.....	.....	.....
Siswa dapat menyebutkan perbedaan antara kedua model	.....	.....	.....	.....
Siswa dapat memberi contoh-contoh yang baik pada dua model	.....	.....	.....	.....
Jawaban sangat baik, jernih, terorganisir, menunjukkan pengertian yang mendalam mengenai kedua model	.....	.....	.....	.....
Jumlah Skor yang Diperoleh				.....

Untuk soal-soal yang sangat terstruktur, seperti pada mata pelajaran Matematika, jawaban yang benar sudah dapat ditebak, sehingga dapat diberikan skor pada rubrik itu jika peserta tes mengerjakan benar suatu langkah pengerjaan.

Contoh 4.4

Diberikan soal berikut.

> Ingin diuji apakah terdapat kesamaan rerata antara kelompok eksperimen dan kelompok kontrol. Banyaknya anggota sampel masing-masing kelompok adalah 18. Rerata sampel kelompok eksperimen adalah 108.10 dengan variansi 289.00. Rerata sampel kelompok kontrol adalah 98.40 dengan variansi 196.00. Diasumsikan variansi-variansi populasinya sama. Bagaimana kesimpulannya?

Maka rubriknya dapat dibuat sebagai berikut. (Bilangan-bilangan di sebelah kanan merupakan skor untuk bagian itu. Jumlah skor adalah 50 jika peserta tes menjawab dengan sempurna)

a. $H_0: \mu_1 = \mu_2$ (rerata kelompok eksperimen sama dengan rerata kelompok kontrol)
   $H_1: \mu_1 \neq \mu_2$ (rerata kelompok eksperimen tidak sama dengan rerata kelompok kontrol) 5

b. Statistik uji yang digunakan: $t = \dfrac{(\overline{X}_1 - \overline{X}_2)}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t(n_1+n_2-2)$ 5

c. Komputasi:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(17)(289) + (17)(196)}{18+18-2} = 24250; \text{.........} \quad 10$$

$$s_p = \sqrt{24250} = 15.572 \text{ ....} \quad 5$$

$$t_{obs} = \frac{108.1 - 98.4}{15.572\sqrt{\frac{1}{18} + \frac{1}{18}}} = 1.87 \text{ ....} \quad 5$$

d. Daerah Kritis:

$t_{0.025;34} = 1.960$; DK = $\{t \mid t < -1.96 \text{ atau } t > 1.96\}$; 5

dan $t_{obs} = 1.87 \notin DK$ 5

e. Keputusan Uji: H_0 diterima. 5

f. Kesimpulan: kelompok eskspcrimen dan kelompok kontrol sama pandainya.

.... 5

Jumlah Skor 50

TES JAWABAN SINGKAT (*SHORT ANSWER TEST*)

Pada tes jawaban singkat, peserta tes diminta untuk memberikan kata, frasa, bilangan, atau simbol untuk menjawab pertanyaan. Tes ini biasanya diberikan di sekolah dasar dan sekolah menengah pertama. Untuk tingkat SMA dan yang sederajad ke atas, tes jawaban singkat ini jarang dipakai.

Ada dua jenis tes jawaban singkat, yaitu: (1) dalam bentuk pertanyaan dan (2) dalam bentuk kalimat yang tidak lengkap. Bentuk kedua ini sering disebut bentuk isian singkat.

Contoh 4.5

Berikut ini contoh dalam bentuk pertanyaan.

1. Berapakah bilangan prima terkecil? ____________________
2. Siapa raja Majapahit pertama? ____________________

Contoh 4.6

Berikut ini contoh dalam bentuk kalimat yang tidak lengkap.

1. Bilangan prima terkecil adalah ____________________
2. Raja Majapahit pertama adalah ____________________

TES MEMILIH JAWABAN (*SELECTED-RESPONSE TEST*)

Tes memilih-jawaban adalah tes yang menghendaki peserta tes untuk memilih di antara kemungkinan-kemungkinan jawaban yang telah disediakan. Tes ini sering disebut tes objektif.

Keunggulan tes memilih-jawaban (*selected-response*), antara lain: (1) mudah, cepat, dan objektif dalam pemberian skor, (2) dapat mencakup bahan yang sangat luas, (3) kemungkinan jawaban yang salah dan yang benar dapat dengan mudah dilihat, dan (4) butir soal dengan ini dapat digunakan berulang kali.

Di sisi lain, kelemahan tes memilih-jawaban (*selected-response*), antara lain: (1) sulit dipakai untuk mengukur aspek kemampuan tingkat tinggi, (2) memerlukan waktu yang lama dalam penyusunan soalnya, (3) jawaban soal tes memilih jawaban dapat diterka, dan (4) tidak dapat membedakan secara jelas kemampuan masing-masing peserta tes.

Berikut ini diberikan saran dalam mengkonstruksi tes memilih-jawaban: (1) usahakan agar kesukaran membaca sesedikit mungkin, (2) jangan semata-mata hanya mengutip dari buku, (3) masing-masing butir soal harus saling independen, tidak saling mempermudah atau mempersulit butir soal yang lain, (4) jika menggunakan lambang-lambang atau simbol-simbol, hendaknya dijelaskan arti lambang-lambang atau simbol-simbol tersebut, (5) dalam menulis soal matematika, hendaknya jangan mengacaukan antara bahasa matematika dan bahasa verbal, (6) hendaknya menggunakan kaidah-kaidah kebahasaan yang benar, dan (8) soal-soal yang telah selesai didraft, hendaknya direview lebih dulu.

Tipe Tes Memilih-Jawaban (*Selected-Response*)

Secara garis besar, tes memilih-jawaban dapat dibedakan atas tiga jenis, yaitu: (1) tes benar-salah, (2) tes menjodohkan, dan (3) tes pilihan ganda (*multiple choice test*). Dari ketiga jenis itu, yang paling sering dipakai (terutama di tingkat SMA/SMK ke atas) adalah jenis tes pilihan ganda. Pada buku ini hanya didiskusikan jenis terakhir.

TES PILIHAN GANDA

Tes pilihan ganda dapat dibedakan atas 9 bentuk, yaitu bentuk: (1) melengkapi lima pilihan, (2) asosiasi dengan lima pilihan, (3) hal kecuali, (4) analisis hubungan antar hal, (5) analisis kasus, (6) perbandingan kuantitatif, (7) hubungan dinamik, (8) melengkapi berganda, dan (9) pemakaian gambar, diagram, dan/atau grafik. Dari berbagai bentuk itu yang paling sering dipakai adalah: (1) bentuk melengkapi lima pilihan, (2) bentuk analisis kasus, dan (3) bentuk melengkapi berganda.

Tes bentuk pilihan ganda terdiri dari batang tubuh yang berupa suatu pernyataan yang belum lengkap atau suatu pertanyaan yang diikuti oleh sejumlah kemungkinan jawaban. Batang tubuh tadi sering disebut pokok soal (*stem*). Kemungkinan jawaban disebut *option*. *Option* yang

merupakan jawaban yang benar disebut kunci (*key*) dan *option-option* yang bukan kunci jawaban disebut pengecoh (distraktor, umpan).

Soal-soal bentuk pilihan ganda lebih fleksibel dan lebih efektif daripada bentuk-bentuk lain. Jika dikonstruksi dengan baik, soal bentuk pilihan ganda amat efektif untuk mengukur kemampuan menguraikan informasi, perbendaharaan kata-kata, aplikasi suatu konsep, atau kemampuan menginterpretasikan sesuatu. Kecuali itu, jika dikonstruksi dengan baik, soal pilihan ganda juga dapat mendiskriminasikan, menentukan pendapat, dan menarik kesimpulan. Satu-satunya kemampuan yang tidak dapat diukur dengan soal tipe pilihan ganda adalah kemampuan mengorganisir sesuatu.

Mengkonstruksi tes pilihan ganda dengan baik sangat sukar dan memerlukan waktu lama. Tidak jarang pembuat soal hanya memasukkan hal-hal yang mudah-mudah saja, yaitu yang sekedar mengukur hal-hal yang bersifat pengetahuan (hafalan)[3].

Berikut ini diberikan beberapa saran jika tes bentuk pilihan ganda ingin digunakan.

(1) Gunakan bahasa Indonesia yang efisien, baik, dan benar. Jangan membuat kalimat terlalu panjang yang dapat membingungkan peserta didik.

(2) Berilah petunjuk pengerjaan yang singkat tetapi jelas. Petunjuk pengerjaan itu harus pula memuat cara memilih alternatif jawaban, misalnya dengan cara menyilang, melingkari, atau menghitami alternatif jawaban yang disediakan.

(3) Alternatif jawaban disusun vertikal ke bawah (tidak ke samping). Hal ini untuk menjamin kenyamanan pandang dan untuk memudahkan para peserta tes untuk melakukan *scanning*.

(4) *Stem* dan seluruh alternatif jawaban harus tercetak pada halaman yang sama. Hal ini untuk memudahkan peserta tes membaca butir soal tersebut.

(5) Jika *stem* merupakan kalimat lengkap yang berupa pertanyaan, tidak perlu diberi tanda baca titik pada akhir alternatif.

Contoh 4.7

Siapakah presiden Rrepublik Indonesia yang pertama?

a. Suharto
b. Sukarno
c. Megawati
d. B.J. Habibie
e. Joko Widodo

---

[3] **Hal ini lah** yang sering **dipakai oleh para** pengritik tes **pilihan ganda untuk** mengumpat tes **pilihan ganda.** Pada pengritik **tidak memahami** bahwa kalau **dikonstruksi** dengan baik, tes **pilihan ganda** dapat mengukur **kemampuan** tingkat tinggi.

(6) Jika *stem* berupa kalimat yang belum lengkap. maka pada akhir *stem* diberi tanda baca tiga titik dan pada akhir alternatif jawaban harus diberi tanda baca titik atau tanda baca titiknya di letakkan pada akhir *stem*.

Contoh 4.8

Presiden Republik Indonesia yang pertama adalah …

a. Suharto.
b. Sukarno.
c. Megawati.
d. B.J. Habibie.
e. Joko Widodo.

(Perhatikanlah bahwa hanya ada tiga titik pada akhir *stem*. Tidak boleh lebih dan tidak boleh pula kurang dari tiga titik)

Contoh 4.9

Presiden Republik Indonesia yang pertama adalah … .

a. Suharto
b. Sukarno
c. Megawati
d. B.J. Habibie
e. Joko Widodo

(7) Jika pada alternatif jawaban memuat satuan ukuran, hendaknya satuan ukuran tersebut diletakkan pada *stem*.

Contoh 4.10

Suatu persegi panjang mempunyai ukuran panjang 10 cm dan lebar 6 cm. Luas persegi panjang tersebut adalah … cm^2.

a. 16
b. 32
c. 60
d. 120
e. 136

(8) *Stem* harus memuat informasi yang lengkap tetapi tidak berlebih dan menanyakan satu ide saja.

Contoh 4.11 (kurang baik)

Suharto … .

a. presiden pertama Republik Indonesia
b. lahir dan meninggal di Jakarta
c. menikah setelah menjadi presiden
d. memerintah hanya satu periode (5 tahun) saja
e. menjabat gubernur sebelum menjadi presiden

(Contoh ini kurang baik karena kurang jelas ide apa yang ditanyakan pada butir soal)

Contoh 4.12 (kurang baik)
Terdapat beberapa skala pengukuran yang digunakan dalam penelitian sosial. Manakah skala pengukuran yang mengakomodasi nilai nol mutlak?
a. interval
b. nominal
c. ordinal
d. rasio
e. likert
(Contoh ini kurang baik karena memuat kalimat pertama yang sebenarnya tidak perlu)

(9) Jika alternatif jawaban berupa bilangan-bilangan, urutkan mulai dari terkecil atau mulai dari terbesar.

(10) Sediakan antara tiga sampai dengan lima alternatif jawaban. Biasanya, di tingkat SD dan SMP disediakan empat alternatif jawaban, sedangkan untuk SMA ke atas disediakan lima alternatif jawaban.

(11) Jika alternatif jawaban berupa kalimat, sediakan kalimat-kalimat yang kira-kira sama panjang.

Contoh 4.13 (kurang baik)
Manakah yang merupakan sifat jajar genjang?
a. mempunyai empat diagonal
b. mempunyai paling sedikit dua sisi yang sejajar dan sama panjang
c. diagonal saling tegak lurus
d. ada dua sisi yang saling tegak lurus
e. diagonal sama panjang

(12) Dihindari menggunakan kata tidak. Jika terpaksa harus digunakan, tulislah kata tidak dengan huruf besar (kapital).

Contoh 4.14
Di Pulau Jawa, propinsi manakah yang TIDAK mempunyai gunung berapi yang masih aktif?
a. Jawa Timur
b. Jawa Tengah
c. Jawa Barat
d. DIY
e. Madura

(13) Pastikan bahwa hanya ada satu jawaban yang benar atau jawaban yang paling tepat.

(14) Jangan menggunakan alternatif jawaban "bukan salah satu di atas" atau "semua benar". Penggunaan "bukan salah satu di atas" dan "semua benar" sebagai alternatif jawaban menunjukkan bahwa pembuat soal tidak mempersiapkan pengecoh dengan baik. Jika terpaksa menggu-

nakan alternatif jawaban "bukan salah satu di atas", gunakanlah sesekali saja (jarang-jarang).

(15) Jangan memasang alternatif jawaban yang jelas-jelas salah atau jelas-jelas benar.

(16) Jika terdapat gambar, diagam, grafik, dan semacamnya, letakkanlah di bagian kiri stem. Sebaiknya jangan meletakkan di sebelah kanan stem.

(17) Jika memuat bilangan desimal, buatlah dengan banyak angka di belakang koma yang sama

Contoh 4.15 (kurang baik)

Panjang sisi suatu persegi adalah 0.5 m. Luas persegi tersebut adalah ... m^2.

a. 0.25
b. 0.5
c. 0.75
d. 1.0
e. 2.00

Contoh 4.16 (lebih baik)

Panjang sisi suatu persegi adalah 0,5 m. Luas persegi tersebut adalah ... m^2.

f. 0,25
g. 0,50
h. 0,75
i. 1,00
j. 2,00

(18) Antara stem dan alternatif jawaban harus gayut (tersambung baik).

Contoh 4.17 (kurang baik)

Jika A adalah himpunan penyelesaian dari $x^2 - 5x + 6 = 0$ maka A = ...

a. akar-akarnya adalah 2 dan 3
b. akar-akarnya adalah −2 dan −3
c. akar-akarnya adalah 1 dan 5
d. akar-akarnya adalah −1 dan −5
e. akar-akarnya adalah 6 dan 11

(19) Butir soal tertentu jangan tergantung kepada butir soal yang lainnya.

Contoh 4.18

Misalnya terdapat dua butir soal (nomor 4 dan 5) berikut ini.

4. Luas suatu persegi adalah $36\,cm^2$. Panjang sisi persegi itu adalah ... cm.
   a. 3
   b. 4
   c. 5
   d. 6
   e. 18

5. Keliling persegi pada soal Nomor 4 adalah ... cm.
   a. 12
   b. 16
   c. 20
   d. 24
   e. 72

(Perhatikanlah bahwa jika seorang peserta tes menjawab salah butir soal nomor 4, maka dia pasti salah menjawab butir soal nomor 5)

(20) Pengecoh harus disusun sama kuat daya tariknya. Untuk membuat pengecoh yang sama kuat daya tariknya, misalnya, dengan membuat semua pengecoh sama panjangnya, sama jenisnya, dan semacamnya.

(21) Jangan menggunakan kata "selalu", "kadang-kadang", dan "tidak pernah" pada alternatif jawaban. Alternatif jawaban yang diawali kata "selalu" atau "tidak pernah" cenderung bukan kunci jawaban. Alternatif jawaban yang diawali dengan kata "kadang-kadang" cenderung merupakan kunci jawaban.

(22) Kalimat-kalimat pada *stem* hendaknya dibuat pendek-pendek untuk memperjelas kalimat.

(23) Karena alternatif jawaban harus disusun vertikal ke bawah, maka untuk menghemat kertas, buatlah susunan soal dalam dua kolom.

(24) Perhatikan baik-baik banyaknya butir soal yang diujikan. Perkirakan seberapa lama peserta tes mengerjakan setiap butirnya. Biasanya, untuk soal-soal Matematika, setiap butir diperkirakan dapat diselesaikan dalam waktu 3 menit, dan untuk soal-soal ilmu pengetahuan sosial, setiap butir soal diperkirakan dapat diselesaikan dalam waktu 2 menit.

(25) Dari sejumlah butir soal yang diujikan, susunlah mulai dari butir soal yang paling mudah (letakkanlah butir-butir yang mudah di awal-awal nomor)

(26) Tempatkan secara random kunci jawaban. Artinya, kunci jawaban jangan diletakkan berpola, misalnya lima butir soal pertama kuncinya a, lima butir kedua kuncinya b, lima butir ketiga kuncinya c, dan seterusnya.

Berikut ini beberapa contoh butir soal bentuk pilihan ganda dalam bentuk Melengkapi Lima Pilihan, Bentuk Analisis Kasus, Bentuk Melengkapi Berganda.

Bentuk Melengkapi Empat atau Lima Pilihan

Bentuk ini adalah bentuk soal pilihan ganda yang banyak digunakan orang. Butir soal pada bentuk ini terdiri dari pokok soal yang diikuti oleh empat atau lima buah alternatif jawaban. Contoh-contoh 4.7 sampai dengan 4.18 di depan adalah contoh butir soal pilihan ganda dalam bentuk melengkapi lima pilihan. Untuk siswa SD dan SMP biasanya hanya ada empat alternatif jawaban.

Contoh 4.19

Berikut ini adalah contoh dua butir soal dalam bentuk melengkapi empat pilihan.

Petunjuk:

Pilihlah salah satu jawaban yang Anda anggap benar di antara kemungkinan kemungkinan jawaban yang benar dengan memberi tanda silang pada lembar jawaban!

1. Luas bayangan ΔPQR dengan $P(1,0)$, $Q(6,0)$, dan $R(6,3)$ oleh transformasi yang sesuai dengan matriks $\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ adalah ... satuan luas.
   a. 15
   b. 30
   c. 45
   d. 60

2. Pernyataan $(p \rightarrow q) \vee r$ bernilai salah, jika ... .
   a. p salah, q salah, dan r benar
   b. p benar, q benar, dan r salah
   c. p salah, q salah, dan r salah
   d. p benar, q salah, dan r salah

Bentuk Analisis Kasus

Butir soal yang dinyatakan dalam bentuk analisis kasus, dimulai dari semacam cerita yang disebut kasus. Dari kasus ini dapat muncul beberapa butir soal yang masing-masing butir soal itu biasanya berbentuk melengkapi empat atau lima pilihan.

Contoh 4.20

Berikut ini adalah butir-butir soal berbentuk analisis kasus pada mata pelajaran Matematika.

Petunjuk:

Ikutilah kasus di bawah ini. Kemudian. pilihlah salah satu jawaban yang Anda anggap benar di antara kemungkinan-kemungkinan jawaban yang benar dengan memberi tanda silang pada lembar jawaban!

Amir dan Ani duduk pada kelas yang sama. Pada semester ini, ia menempuh 15 mata pelajaran. Kemarin, mereka menerima buku rapor mereka. Nilai-nilai mereka tampak seperti berikut ini.

Amir : 7, 9, 10, 7, 8, 8, 9, 8, 8, 8, 8, 6, 7, 8, 9
Ani : 8, 8, 8, 8, 9, 8, 9, 8, 7, 8, 8, 9, 8, 7, 7

1. Rerata nilai Ani adalah ... .
   a. 5
   b. 6
   c. 7
   d. 8
   e. 9

2. Jangkauan nilai Amir adalah ... .
   a. dua kali jangkauan nilai Ani
   b. satu lebihnya dari jangkauan nilai Ani
   c. sama dengan jangkauan nilai Ani
   d. sama dengan nol
   e. setengah kali jangkauan nilai Ani

Bentuk Melengkapi Berganda

Kalau pada bentuk melengkapi lima pilihan, hanya terdapat satu jawaban yang benar (atau paling tepat), pada bentuk ini terdapat beberapa jawaban yang benar, tetapi untuk menjawab butir soal tersebut, ada beberapa kombinasi di antaranya.

Contoh 4.21

Berikut ini adalah contoh butir soal dalam bentuk melengkapi berganda.

Petunjuk:

Di bawah ini terdapat butir-butir soal yang mempunyai kejadian yang dapat muncul bersama-sama. Pada lembar jawaban, silanglah:

a. jika hanya 1, 2, dan 3 yang benar
b. jika hanya 1 dan 3 yang benar
c. jika hanya 2 dan 4 yang benar
d. jika hanya 4 yang benar
e. jika 1, 2, 3, dan 4 benar

1. Yang merupakan himpunan kosong adalah ... .
   1. Himpunan dari semua himpunan
   2. Himpunan bilangan genap yang habis dibagi dua
   3. Himpunan bilangan cacah yang kurang dari 10
   4. Himpunan yang anggotanya bilangan asli yang terbesar

2. Jika $y = x^3 + x^2 + 5$, maka ... .
   1. $y(0) = 5$
   2. $y(1) = 7$
   3. $y(2) = 14$
   4. $y(3) = 41$

TAKSONOMI BLOOM

Seperti disebutkan di Pendahuluan, terdapat tiga ranah tujuan pembelajaran, yaitu: (a) tujuan pada ranah kognitif, (b) tujuan pada ranah afektif, dan (c) tujuan pada ranah psikomotor. Tes hasil belajar, seharusnya mengukur kemampuan pada ketiga ranah tersebut. Namun demikian, ada mata-mata pelajaran tertentu yang yang lebih berat ke ranah tertentu. Ujian pada mata-pelajaran Matematika, misalnya, lebih bersifat mengukur kemampuan pada ranah kognitif daripada ranah psikomotor. Ujian praktik pada mata-pelajaran Seni Suara, lebih mengukur kemampuan di ranah psikomotor daripada ranah kognitif.

Terdapat banyak penggolongan tujuan pembelajaran pada ranah kognitif, salah satu di antaranya adalah penggolongan tujuan pembelajaran berdasarkan taksonomi Bloom. Menurut Bloom, tujuan pembelajaran pada ranah kognitif pada dasarnya dapat dibedakan menjadi 6 tingkatan hierarkis, yaitu: (1) pengetahuan (*knowledge*, C1), (2) pemahaman (*comprehension*, C2), (3) penerapan (*application*, C3), (4) analisis (*analysis*, C4), (5) sintesis (*synthesis*, C5), dan (6) evaluasi (*evaluation*, *C6*).

Aspek Pengetahuan

Tujuan pembelajaran pada aspek pengetahuan berkenaan dengan ingatan bahan yang telah dipelajari, yang biasanya cenderung bersifat hafalan. Tujuan pada aspek ini telah tercapai apabila siswa sudah mampu menyebutkan kembali informasi yang telah diperolehnya. Tujuan pada aspek ini sudah dapat diungkap apabila siswa telah ingat dan dapat menyebutkan tentang: simbul, fakta, konsep, definisi, dalil, klasifikasi, terminologi dan semacamnya.

Contoh 4.22

Berikut ini adalah contoh butir soal pada aspek pengetahuan dalam bentuk melengkapi lima pilihan.

Pilihlah salah satu jawaban yang Anda anggap benar di antara kemungkinan-kemungkinan jawaban yang benar dengan memberi tanda silang pada lembar jawaban!

1. Formula yang ditulis dalam bentuk $a^2 + b^2 = c^2$ disebut formula ... .
   a. Pythagoras
   b. Euler
   c. Archimides
   d. De'l Hospital
   e. Fibbonaci

2. Lambang $\frac{4}{5}$ adalah lambang untuk ... .
   a. bilangan asli
   b. bilangan cacah
   c. pecahan
   d. bilangan kompleks
   e. bilangan bulat

Aspek Pemahaman

Tujuan pembelajaran pada aspek pemahaman berkenaan dengan kemampuan memahami arti suatu bahan pelajaran, namun dalam tingkatan yang rendah, misalnya mampu mengubah suatu informasi ke dalam informasi lain yang lebih bermakna dan memberikan suatu interpretasi. Perbuatannya itu dilakukan atas suruhan tanpa ada kaitannya dengan yang lain. Juga tidak dituntut pemakaiannya dalam situasi yang lain.

Menurut Bloom, tujuan pada aspek pemahaman dapat dibedakan menjadi tiga bagian, yaitu: (a) pengubahan (*translation*), (b) pemberian arti (*interpretation*), dan (c) pemerkiraan (*extrapolation*).

Contoh 4.23

Berikut ini adalah contoh butir soal pada mata pelajaran Matematika aspek pemahaman dalam bentuk melengkapi lima pilihan.

Pilihlah salah satu jawaban yang Anda anggap benar di antara kemungkinan-kemungkinan jawaban yang benar dengan memberi tanda silang pada lembar jawaban!

1. Bilangan 100010_{dua} dapat ditulis sebagai ... .
   a. 100010
   b. 48
   c. 45
   d. 42
   e. 34

2. Himpunan penyelesaian dari $x^2 - 4 = 0$ adalah ... .
   a. Φ
   b. $\{4\}$
   c. $\{2\}$
   d. $\{-2, 2\}$
   e. $\{-4, 4\}$

Aspek Penerapan

Tujuan pembelajaran pada aspek penerapan berkenaan dengan penggunaan ketentuan-ketentuan, prinsip-prinsip, dan/atau konsep-konsep

yang telah diterima siswa. Tujuan pada aspek ini telah tercapai jika siswa telah dapat menggunakan apa yang telah diperolehnya dalam situasi khusus yang baru, baik yang masih terdapat dalam satu mata pelajaran maupun penggunaannya di mata pelajaran lain.

Contoh 4.24

Berikut ini adalah contoh butir soal untuk mata pelajaran Matematika pada aspek penerapan.

Pilihlah salah satu jawaban yang Anda anggap benar di antara kemungkinan-kemungkinan jawaban yang benar dengan memberi tanda silang pada lembar jawaban!

1. Aku adalah suatu bilangan. Jika aku dikalikan 7 dan kemudian ditambah dengan kuadrat aku, maka hasilnya adalah nol. Andaikan aku adalah bilangan bulat, maka aku adalah ... .
   a. 0
   b. 7
   c. −7
   d. −7 atau 0
   e. 0 atau 7

2. Misalnya terdapat papan catur raksasa. Seseorang meletakkan 1 butir jagung pada kotak ke-1 papan catur tersebut, 2 butir jagung pada kotak ke-2, 4 butir jagung pada kotak ke-3, 8 butir jagung pada kotak ke-4, dan seterusnya dengan menggunakan aturan yang sama. Banyaknya butir jagung pada kotak terakhir papan catur tersebut adalah ... .
   a. tak dapat dihitung
   b. 2^{61} butir
   c. 2^{62} butir
   d. 2^{63} butir
   e. tak terhingga butir

Aspek Analisis

Tujuan pembelajaran pada aspek analisis ingin melihat apakah siswa telah dapat mengurai suatu sistem ke dalam bagian-bagiannya, mencari hubungan antara bagian-bagiannya, dan mengenal bagian-bagian itu sebagai satu sistem yang baru.

Contoh 4.25

Berikut ini adalah contoh butir soal untuk matapalajaran Matematika pada aspek analisis.

Pilihlah salah satu jawaban yang Anda anggap benar di antara kemungkinan-kemungkinan jawaban yang benar dengan memberi tanda silang pada lembar jawaban!

1. Diketahui m dan n bilangan ganjil positif yang kurang daripada 5 dengan n < m. Bilangan genap positif terbesar yang dapat membagi bilangan dengan bentuk $m^2 - n^2$ adalah ....
   a. 2
   b. 4
   c. 6
   d. 8
   e. 10

2. Setiap bilangan rasional mempunyai invers perkalian, kecuali ....
   a. 1
   b. 0
   c. −1
   d. 2
   e. −2

Aspek Sintesis

Tujuan pembelajaran pada aspek sintesis ingin melihat apakah siswa telah dapat bekerja dengan bagian-bagian, elemen-elemen, atau unsur-unsur untuk kemudian menyusunnya menjadi suatu sistem yang baru. Jadi, aspek sintesis berkaitan dengan kemampuan seseorang untuk menyusun sesuatu yang baru dari berbagai unsur, konsep, pola, aturan, dan sebagainya. Unsur-unsur yang telah ia miliki harus ia organisasikan untuk memperoleh sesuatu yang baru.

Menulis soal dalam mata pelajaran tertentu, misalnya Matematika, pada aspek sintesis biasanya sangat sukar, karena sifat matematika yang bersifat terstruktur. Apalagi kalau bentuk butir soalnya adalah pilihan ganda.

Contoh 4.26

Berikut ini adalah contoh butir soal pada mata pelajaran Matematika pada aspek sintesis pada bentuk soal uraian.

A. Kerjakan soal-soal berikut ini.

1. Buktikan bahwa jumlah n bilangan ganjil yang pertama adalah n^2.
2. Tunjukkan bahwa $A = \{x \mid x^3 = 1\}$ adalah grup pada operasi perkalian.

B. Pilihlah salah satu jawaban yang Anda anggap benar di antara kemungkinan-kemungkinan jawaban yang benar dengan memberi tanda silang pada lembar jawaban!

1. Jika A adalah himpunan penyelesaian dari $x^3 - 2x^2 + x = 0$. maka banyaknya himpunan bagian A adalah ... .
   a. 0
   b. 1
   c. 2
   d. 4
   e. 8

2. Jika p dan q adalah akar-akar dari $x^2 - x - 12 = 0$ dan $p < q$, maka $\int_p^q (2x + 4)dx = \ldots$.
   a. 28
   b. 30
   c. 35
   d. 40
   e. $2p + 4q$

Aspek Evaluasi

Tujuan pembelajaran pada aspek evaluasi telah dapat dicapai oleh siswa jika siswa telah mampu membuat kriteria. memberikan pertimbangan. mengkaji (kekeliruan, ketepatan, ketetapan), dan mampu menilai. Aspek evaluasi merupakan aspek kelompok kognitif tertinggi tingkatannya, sebab menyangkut semua aspek yang lain.

Menulis butir soal dalam mata pelajaran tertentu, misalnya mata pelajaran Matematika, pada aspek evaluasi biasanya juga sangat sukar. Menulis butir soal untuk mengukur aspek evaluasi dengan bentuk pilihan ganda juga sangat sukar.

Contoh 4.27

Berikut ini adalah contoh soal pada aspek evaluasi.

Jawablah soal-soal berikut ini.

1. Beberapa orang mengatakan bahwa sistem desimal adalah sistem penulisan bilangan yang paling unggul dibandingkan dengan sistem yang lain, misalnya sistem penulisan bilangan dengan cara Romawi. Jelaskan mengapa orang berpendapat seperti itu!
2. Dua dari banyak permasalahan di kota besar adalah peledakan penduduk dan kemacetan lalu lintas. Buatlah perencanaan kota yang dapat mengatasi kemacetan lalu lintas, namun tetap nyaman bagi lingkungan padat penduduk.

Jika kita membuat tes yang mengungkap aspek pengetahuan (C1) dan pemahaman (C2) saja. berarti kita hanya ingin mengukur kemampuan tingkat rendah. Sebaliknya, jika kita membuat tes yang mengungkap aspek penerapan (C3), analisis (C4), sintesis (C5), dan evaluasi (C6), maka berarti kita mengukur kemampuan tingkat tinggi (*higher order thingking*).

Perlu diketahui bahwa batas antara aspek yang satu dengan aspek yang lain tidak dapat dibuat definitif, sehingga kadang-kadang agak sukar membedakan ciri-ciri soal yang mengungkap masing-masing aspek. Juga tidak semua bentuk tes cocok untuk mengungkap tujuan di semua aspek. Tes pilihan ganda, misalnya agak sukar mengungkap tujuan-tujuan di aspek sintesis dan evaluasi, tetapi sangat mudah dipakai untuk mengungkap tujuan-tujuan di aspek pengetahuan, pemahaman, dan penerapan.

TAKSONOMI BLOOM YANG DIREVISI

Anderson dan Krathwol (2001: 67-68) mengemukakan bahwa dimensi dari proses kognitif dibedakan menjadi 6 tingkatan[4], yaitu: (1) *remember* (mengingat), (2) *understand* (mengerti), (3) *apply* (menggunakan), (4) *analyse* (menganalisis), (5) *evaluate* (mengevaluasi), dan (6) *create* (membentuk). Penjelasan masing-masing tingkatan diuraikan secara singkat berikut.

Remember (Mengingat)

Kegiatan pembelajaran disebut pada tingkatan *remember* (mengingat) jika seseorang dapat *retrieve relevant knowledge from long-term memory* (mengungkap kembali pengetahuan yang relevan dari memori jangka panjang).

Tingkatan ini terbagi menjadi: (1) *recognizing*, yaitu mengidentifikasi pengetahuan pada memori jangka panjang yang cocok dengan materi yang disajikan, misalnya mengidentifikasi hari-hari penting dalam sejarah kemerdekaan RI, dan (2) *recalling*, yaitu memanggil kembali pengetahuan dari memori jangka panjang, misalnya mengingat kembali peristiwa penting dalam sejarah kemerdekaan RI.

Understand (Mengerti)

Kegiatan pembelajaran disebut pada tingkatan *understand* (mengerti) jika seseorang dapat *construct meaning from instructional messages, including oral, writen and graphic communications* (membentuk arti dari pesan pembelajaran, termasuk pembelajaran lisan, tertulis, atau memalui komunikasi gambar.

Tingkatan ini terbagi menjadi: (1) *interpreting* (misalnya menyatakan bentuk numerik ke bentuk verbal; menarasikan percakapan dari dokumen penting); (2) *exempliying* (memberi contoh atau ilustrasi khusus dari

[4] Taksonomi ini sering disebut revisian dari Taksonomi Bloom menurut Anderson dan Krathwol.

suatu konsep atau prinsip, misalnya memberi contoh berbagai model lukisan artistik); (3) *classifying* (menentukan bahwa sesuatu termasuk atau tidak termasuk suatu kelompok, misalnya mengklasifikasikan kasus-kasus *mental disorders*); (4) *summarizing* (menyimpulkan point-point penting dari suatu diskusi atau bacaan atau yang lain); (5) *infering* (menarik kesimpulan logis dari informasi yang ada); (6) *comparing* (mendeteksi hubungan antara dua ide, objek, dan semacamnya); dan (7) *explaining* (membangun model sebab akibat dari suatu sistem).

***Apply* (Menggunakan)**

Kegiatan pembelajaran disebut pada tingkatan menggunakan jika seseorang dapat *carry out or use procedure in a given situation*. Tingkatan ini terbagi menjadi: (1) *executing* (menggunakan suatu prosedur untuk mengerjakan tugas tertentu yang sudah familier, misalnya membagi sebuah bilangan dengan bilangan lain, yang masing-masing bilangan terdiri dari 3 digit), dan (2) *implementing* (menggunakan suatu prosedur untuk mengerjakan tugas tertentu yang belum familier).

***Analyze* (Menganalisis)**

Kegiatan pembelajaran disebut pada tingkatan menganalisis jika seseorang dapat *breaks material into its constituent parts and determine how the part related to one another and to overall structure or purpose*. Tingkatan ini terbagi menjadi: (1) *differentiating* (yaitu membedakan bagian yang relevan dari bagian yang tidak relevan atau membedakan bagian yang penting dari bagian yang tidak penting dari suatu material tertentu, misalnya membedakan antara bilangan yang relevan dan tidak relevan pada suatu soal cerita), (2) *organizing* (yaitu menentukan bagaimana suatu elemen cocok atau berfungsi dalam suatu struktur atau organisasi), dan (3) *attributing* (menentukan *point of view, bias, values, or intent* pada suatu materi yang disajikan, misalnya sebutkan pokok-pokok penting dari pengarang pada tulisannya dilihat dari perspektif politik dewasa ini).

***Evaluate* (Menilai)**

Kegiatan pembelajaran disebut pada tingkatan menilai jika seseorang dapat *make judgements based on criteria and standards*. Tingkatan ini terbagi menjadi: (1) *checking* (yaitu mendeteksi kesalahan atau kekeliruan pada suatu proses atau produk, menentukan apakah suatu proses atau produk mempunyai konsistensi, atau menentukan keefektifan suatu prosedur yang sedang dilakukan, dan (2) *critiquing* (yaitu mendeteksi ketidakkonsistenan antara produk dan kriteria eksternal, mendeteksi apakah suatu produk

konsisten dengan kriteria luar yang ditentukan, atau mendeteksi ketepatan suatu prosedur untuk permasalahan tertentu, misalnya menjustifikasi di antara dua metode, manakah yang paling baik untk menyelesaikan permasalahan yang ditentukan.

***Create* (Membentuk)**

Kegiatan pembelajaran disebut pada tingkatan membentuk jika seseorang dapat *put elements together to form a coherent or functional whole; reorganize elements into a new pattern or structure.* Tingkatan ini terbagi menjadi: (1) *generating* (yaitu dapat menyebutkan hipotesis alternatif berdasar suatu kriteria), (2) *planning* (menciptakan suatu prosedur untuk mengerjakan tugas-tugas), dan (3) *producing* (yaitu menciptakan sebuah produk).

LANGKAH-LANGKAH KONSTRUKSI TES HASIL BELAJAR PADA RANAH KOGNITIF

Telah diketahui bahwa melakukan penilaian adalah kegiatan rutin seorang guru yang tidak dapat dipisahkan dari kegiatan belajar. Hasil penilaian itu sendiri sangat berguna untuk berbagai pengambilan keputusan mengenai siswa.

Agar keputusan-keputusan yang diambil merupakan keputusan yang bijaksana maka informasi yang dikumpulkan harus benar-benar baik. Untuk memperoleh informasi yang baik, maka alat pengambil informasinya harus benar-benar baik.

Secara garis besar, untuk menyusun tes yang baik, diperlukan langkah-langkah berikut: (1) menginventarisasi bahan yang telah diajarkan, (2) menyusun spesifikasi tes, (3) menyusun butir-butir soal beserta kuncinya, (4) menelaah butir-butir tes, (5) melakukan uji coba, (6) melakukan analisis tes dan analisis butir soal berdasarkan hasil uji coba, (7) melakukan revisi terhadap butir-butir soal yang kurang baik, jika memungkinkan untuk melakukan uji coba lagi, (8) menetapkan instrumen (yang terdiri dari butir-butir yang baik), (9) melaksanakan pengukuran (pengujian) kepada subjek yang dikehendaki, dan (10) menafsirkan hasil yang diperoleh.

Penyusunan Spesifikasi Tes

Penyusunan spesifikasi tes, biasanya, mencakup: penentuan tujuan, pembuatan kisi-kisi, pemilihan jenis tes, dan penentuan banyaknya butir pada setiap kompetensi dasar atau setiap indikator. Kisi-kisi tes, biasanya, ditampilkan dalam bentuk matriks yang menunjukkan isi pokok bahasan

(atau kompetensi dasar) yang akan diukur dan aspek tingkah laku yang akan diungkap.

Untuk spesifikasi tes hasil belajar yang tidak memungkinkan adanya uji coba untuk memperoleh butir-butir yang baik, kadang-kadang disertakan pula level tingkat kesulitan butir soal, apakah termasuk ke dalam kategori mudah, sedang, atau sukar. Beberapa pakar mengatakan bahwa komposisi tingkat kesukaran perangkat tes adalah 25% mudah, 50% sedang, dan 25% sukar.

Contoh 4.28

Berikut ini adalah contoh kisi-kisi untuk tes bentuk pilihan ganda pada suatu ujian, misalnya ujian akhir semester, yang menyertakan level tingkat kesulitan soal.

Mata Pelajaran :
Tahun Ajaran :
Semester :
Lama Ujian :
Banyaknya Butir Soal yang Diperlukan/Diujicobakan:

No	Pokok Bahasan/ Kompetensi Dasar/ Indikator	Jenjang Kemampuan dan Tingkat Kesulitan												Banyaknya Butir Soal	Persentase
		C1			C2			C3			C4, C5, C6				
		Mudah	Sedang	Sukar	Mudah	Sedang	Sukar	Mudah	Sedang	Sukar	Mudah	Sedang	Sukar		
1															
2															
3															
4															
...															
N															
Banyaknya Butir Soal															
Persentase															

Pada umumnya, kisi-kisi untuk soal tipe uraian lebih sederhana, karena pemilahan jenjang berpikir peserta tes menjadi C1, C2, C3, C4, C5, dan C6 tidak perlu diberikan.

Contoh 4.29

Berikut ini adalah contoh kisi-kisi untuk tes bentuk uraian.

Mata Pelajaran :
Tahun Ajaran :
Semester :
Lama Ujian :
Banyaknya Butir Soal yang Diperlukan/Diujicobakan:

No	Pokok Bahasan/ Kompetensi Dasar/ Indikator	Jenis Soal		Banyaknya Butir Soal	Persentase
		Terbatas	Bebas		
1					
2					
3					
4					
...					
N					
Banyaknya Butir Soal					
Persentase					

Jika kisi-kisi dibuat untuk keperluan uji coba dalam suatu penelitian, maka banyaknya butir soal yang akan dipakai untuk uji coba harus lebih banyak dibandingkan dengan banyaknya butir soal yang akan digunakan. Misalnya, untuk ujian dalam waktu 90 menit diperlukan 30 butir soal pilihan ganda. Maka untuk uji coba, diperlukan 35 – 40 butir soal dengan waktu uji coba 120 menit.

Perlu pula diketahui bahwa ada perbedaan mendasar kisi-kisi untuk tes hasil belajar, misalnya pada ujian akhir semester, dengan kisi-kisi tes prestasi belajar untuk suatu penelitian. Untuk kepentingan penelitian diharuskan diperolehnya nilai yang menyebar menurut distribusi normal. Oleh karena itu, pada kisi-kisi tes untuk suatu penelitian, pembagian butir soal menjadi mudah, sedang, dan sukar menjadi tidak relevan. Hal ini disebab-

kan, agar diperoleh nilai yang menyebar, maka butir soal harus berkategori sedang. Pemerolehan kategori sedang tersebut diketahui setelah dilakukan analisis butir soal setelah uji coba.

Contoh 4.30

Berikut ini adalah contoh kisi-kisi untuk tes bentuk pilihan ganda pada suatu uji coba penelitian.

Nama Variabel :
Lama Ujian :
Banyaknya Butir Soal yang Diujicobakan:

No	Pokok Bahasan/ Kompetensi Dasar/ Indikator	Jenis Kemampuan yang Diukur				Banyak-nya Butir Soal	Persentase
		C1	C2	C3	C4. C5. C6		
1							
2							
3							
4							
...							
N							
Banyaknya Butir Soal							
Persentase							

BAHAN DISKUSI

1. Ada orang yang membedakan tiga jenis kemampuan yang perlu diolah pada suatu pembelajaran, yaitu olah pikir, olah rasa, dan olah raga. Apakah menurut Anda taksonomi seperti itu menyerupai taksonomi Bloom? Mengapa?

2. Buatlah kisi-kisi untuk membuat perangkat tes untuk semester tertentu di SMP, SMA, atau SMK yang diujikan dalam waktu 120 menit dalam bentuk pilihan ganda melengkapi lima pilihan.

3. a. Sebutkan hal-hal yang merupakan keunggulan tes uraian!
   b. Sebutkan hal-hal yang merupakan keunggulan tes pilihan ganda!

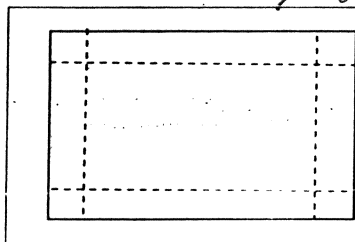
c. Manakah yang lebih unggul, tes bentuk uraian atau tes bentuk pilihan ganda? Jelaskan!

4. Buatlah butir soal pilihan ganda bentuk melengkapi lima pilihan pada bidang studi Anda yang mengungkap aspek:
   a. pengetahuan
   b. pemahaman
   c. aplikasi
   d. sintesis
   e. analisis
   f. evaluasi
   g. membentuk (*create*)

5. Perhatikan butir soal berikut. Butir soal itu mengukur kemampuan siswa mengenai turunan.

Jika $f(x) = (4 - 3x)^2$, maka $f'(x) = \cdots$.
a. $4 - 3x$
b. $18x - 24$
c. $8x + 24$
d. $8 - 6x$
e. $16 - 12x$

Apakah butir soal tersebut mengungkap aspek pengetahuan, pemahaman, atau aplikasi? Mengapa?

6. Perhatikan soal berikut. Butir soal tersebut mengukur kemampuan siswa mengenai turunan.

Sebuah kotak dibuat dari selembar kertas, yang berbentuk persegi panjang yang panjangnya 24 cm dan lebarnya 9 cm, dengan memotong persegi identik pada keempat pojoknya dan melipat ke atas sisi-sisinya. Carilah ukuran kotak agar volumnya maksimum!

Apakah butir soal tersebut mengungkap aspek pengetahuan, pemahaman, atau aplikasi? Mengapa?

7. Perhatikan butir soal berikut. Butir soal tersebut mengungkap kemampuan siswa dalam mencari invers matriks.

Jika $A = \begin{pmatrix} 3 & 5 \\ 4 & 9 \end{pmatrix}$ maka A^{-1} = ...

a. $\begin{pmatrix} 3 & 5 \\ 4 & 9 \end{pmatrix}$

b. $\begin{pmatrix} 9 & -5 \\ -4 & 3 \end{pmatrix}$

c. $\frac{1}{47}\begin{pmatrix} 9 & -5 \\ -4 & 3 \end{pmatrix}$

d. $\frac{1}{7}\begin{pmatrix} 9 & -5 \\ -4 & 3 \end{pmatrix}$

e. bukan salah satu di atas

a. Terkait taksonomi Bloom, mengungkap aspek apakah butir soal tersebut? Mengapa?
b. Berbentuk apakah butir soal tersebut?
c. Apakah butir soal tersebut merupakan butir soal yang baik? Mengapa?

8. Perhatikan butir soal berikut. Butir soal terebut mengungkap kemampuan siswa pada materi logika.

Negasi pernyataan: "jika ada asap, maka ada api" adalah .....

a. jika tak ada api, maka tak ada asap
b. jika ada asap, maka tak ada api
c. ada asap, tetapi tak ada api
d. ada asap atau ada api
e. ada asap atau tak ada api

a. Terkait taksonomi Bloom, mengungkap aspek apakah butir soal tersebut? Mengapa?
d. Berbentuk apakah butir soal tersebut?
e. Dari sisi *lay-out*, apakah butir soal tersebut merupakan butir soal yang *lay-out*nya baik? Mengapa?

9. Perhatikah butir soal berikut.

> Bagaimana menurut pendapat Anda, apakah perkembangan pembelajaran matematika di Indonesia sudah cukup baik atau masih tertinggal dibandingkan dengan perkembangan pembelajaran matematika di Malaysia? Berikan alasan-alasan yang mendukung pendapat Anda tersebut.

   a. Terkait taksonomi Bloom, mengungkap aspek apakah butir soal tersebut? Mengapa?
   b. Berbentuk apakah butir soal tersebut?
   c. Ubahlah butir soal tersebut ke dalam bentuk melengkapi lima pilihan.

10. Perhatikan butir soal beriut. Butir soal itu mengungkap kemampuan siswa dalam menyelesaikan soal program linear.

> **Rokok A yang harga belinya Rp10.000,00 dijual dengan harga Rp11.00,00 per bungkus, sedangkan rokok B yang harga belinya Rp15.00,00 dijual dengan harga Rp17.00,00 per bungkus. Seorang pedagang rokok yang mempunyai modal Rp3.000.000,00 dan kiosnya dapat menampung paling banyak 250 bungkus rokok akan mendapat keuntungan maksimum jika ia membeli ....**
>
> **a. 150 bungkus rokok A dan 100 bungkus rokok B**
> **b. 100 bungkus rokok A dan 150 bungkus rokok B**
> **c. 250 bungkus rokok A dan 200 bungkus rokok B**
> **d. 250 bungkus rokok A saja**
> **e. 200 bungkus rokok B saja**

   a. **Terkait taksonomi Bloom, butir soal itu mengungkap aspek pemahaman atau aplikasi? Mengapa?**
   b. **Adakah kunci jawaban pada butir soal tersebut? Jika ada, yang mana?**
   c. Apakah butir soal tersebut sudah merupakan butir soal yang baik? Mengapa?

11. Perhatikan butir soal berikut.

> Ketika Anda membuat RPP (Rencana Pelaksanaan Pembelajaran) untuk pelaksanaan pembelajaran selama 2 kali 45 menit, Anda pasti ingin mengetahui apakah tujuan pembelajaran telah tercapai atau belum. Penilaian tipe apakah yang baik dituliskan pada RPP tersebut?
> A. Tipe yang baik adalah tipe uraian, sebab mudah membuatnya.
> B. Tipe yang baik adalah tipe uraian, sebab tidak perlu menulis banyak butir soal.
> C. Tipe yang baik adalah tipe pilihan ganda, sebab dapat diskor dengan mudah.
> D. Tipe yang baik adalah tipe pilihan ganda, sebab dapat meliput bahan yang sangat luas.
> E. Bentuk uraian atau bentuk pilihan ganda dapat dipilih, sebab keduanya sama baiknya jika dikonstruksi dengan baik.

a. Terkait taksonomi Bloom, butir soal itu mengungkap aspek yag mana? Mengapa?
b. Adakah kunci jawaban pada butir soal tersebut? Jika ada, yang mana?
c. Apakah butir soal tersebut sudah merupakan butir soal yang baik? Mengapa?

****

lebih baik soal ini disampaikan dlm bentuk soal uraian, shg dpt mengukur sejauh mana

BAB V
ANALISIS BUTIR SOAL PENILAIAN RANAH KOGNITIF

PENDAHULUAN

Untuk kepentingan penelitian atau untuk mendapatkan distribusi skor (nilai) yang menyebar, sebelum tes (soal) digunakan soal-soal tersebut harus diujicobakan terlebih dulu. Dari sisi instrumen, harus dilihat apakah tes telah memenuhi persyaratan validitas atau belum. Dari sisi butir instrumen, butir-butir soal harus dilihat apakah telah memenuhi kelayakan sebagai butir yang baik atau belum. Oleh karena itu, diperlukan analisis butir soal.

ANALISIS BUTIR SOAL UNTUK SOAL PILIHAN GANDA

Pada suatu uji coba, perlu dilihat kualitas butir soal. Kualitas butir soal ditandai oleh tingkat kesulitannya, daya pembedanya, dan berfungsinya pengecoh, jika bentuk soalnya adalah pilihan ganda. Berikut ini diberikan uraian mengenai analisis butir soal untuk tes bentuk pilihan ganda.

TINGKAT KESULITAN (*DIFFICULTY*)

Tingkat kesulitan atau tingkat kesukaran butir soal menyatakan proporsi banyaknya peserta yang menjawab benar butir soal tersebut terhadap seluruh peserta tes. Indeks tingkat kesulitan butir soal dapat dirumuskan dengan rumus berikut.

$$P = \frac{B}{N} \qquad 5.1$$

dengan P adalah indeks tingkat kesulitan butir soal, B adalah banyaknya peserta tes yang menjawab benar butir soal tersebut, dan N adalah banyaknya seluruh peserta tes.

Berdasarkan rumus yang ditulis pada Persamaan 5.1 tersebut, dapat dibuktikan bahwa rentang nilai indeks tingkat kesulitan adalah:

$$0 \leq P \leq 1$$

Pada suatu butir tertentu, nilai P = 0 diperoleh ketika tidak ada satupun peserta tes yang menjawab benar butir itu dan nilai P = 1 diperoleh ketika semua peserta tes menjawab benar butir itu.

Berdasarkan rumus itu pula dapat disimpulkan bahwa semakin tinggi nilai P, maka semakin mudah suatu butir soal dan semakin rendah nilai P maka semakin sukar butir soal tersebut.

Tentu saja, nilai tingkat kesulitan suatu butir tergantung kepada kelompok peserta tes yang dikenai uji coba. Jika kelompok itu merupakan kelompok yang terdiri dari siswa-siswa pandai, maka suatu butir soal cenderung mempunyai tingkat kesulitan yang tinggi (dalam arti butir soal tampak mudah). Di sisi lain, jika kelompok peserta tes yang dikenai uji coba terdiri dari siswa-siwa yang tidak pandai, maka suatu butir soal cenderung mempunyai tingkat kesulitan yang rendah (dalam arti butir soal tampak sulit). Oleh karena itu lah, untuk memperoleh parameter butir soal yang stabil diperlukan sampel uji coba yang cukup besar. Semakin besar ukuran sampel uji coba semakin baik.

Indeks Tingkat Kesulitan yang Diperbolehkan

Pada analisis tingkat kesulitan, pengembang tes harus menentukan kapan suatu butir dipertahankan dalam suatu tes, dibuang, atau direvisi[1].

Dalam konteks penelitian atau penilaian yang menggunakan pendekatan acuan norma (PAN), untuk memperoleh variabel terikat yang semakin menyebar, maka butir soal yang semakin mendekati tingkat kesulitan 0,5, semakin baik. Misalnya peneliti memutuskan bahwa suatu butir soal dipakai jika mempunyai tingkat kesulitan pada interval $0{,}20 \leq P \leq 0{,}80$ atau $0{,}25 \leq P \leq 0{,}75$ atau $0{,}30 \leq P \leq 0{,}70$ tergantung kepada urgensi penelitian. Biasanya, dilihat dari sisi tingkat kesulitan, yang dipakai sebagai kriteria butir yang baik adalah $0{,}30 \leq P \leq 0{,}70$.

[1] Kalau pengembang tes tidak ingin melakukan uji coba berkali-kali untuk suatu tes maka hanya ada dua alternatif penyimpulan, yaitu dipertahankan dalam tes (dipakai) atau dibuang. Alternatif ini lah yang biasanya ditempuh oleh mahasiswa ketika menulis tugas akhirnya.

Pada penilaian yang menggunakan pendekatan acuan patokan (PAP), misalnya pada pelaksanaan kurikulum berbasis kompetensi (KTSP), analisis tingkat kesulitan menjadi tidak relevan untuk dibicarakan, karena yang terpenting pada kurikulum berbasis kompetensi adalah apakah peserta didik telah memenuhi standar minimal kelulusan atau belum. Pada pelaksanaan KTSP, seorang guru pasti mengharapkan semua butir soal dapat dikerjakan oleh semua siswa, yang berarti kalau dikaitkan dengan tingkat kesulitan, maka yang diharapkan adalah butir soal yang tingkat kesulitannya tinggi.

Tentu saja penentuan butir yang baik menurut tingkat kesulitannya bervariasi menurut kepentingannya. Jika tujuan tes adalah untuk *mastery learning*, maka diinginkan butir soal mempunyai indeks tingkat kesulitan sekitar 0,90 (ini berarti diharapkan sekitar 90% siswa akan mencapai tingkat tuntas (*master*)). Jika tujuan tes adalah untuk seleksi, di mana hanya akan diterima 25% pelamar, maka butir soal yang baik untuk itu adalah butir soal yang indeks tingkat kesulitan sekitar 0,25.

Contoh 5.1

Suatu tes pilihan ganda, terdiri dari 15 butir, dikenakan kepada 10 siswa. Sebaran skor untuk masing-masing butir dan skor total peserta tes tampak pada tabel berikut.

Tabel 5.1. Sebaran Skor 10 Siswa dalam Menjawab 15 Butir Soal

Nomor Urut Siswa	Nomor Butir Soal															Skor Total
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	1	1	1	0	1	0	0	1	1	1	0	1	1	0	1	10
2	0	0	1	1	0	0	0	1	1	0	1	0	0	1	0	6
3	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	14
4	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	11
5	0	1	0	0	0	1	1	1	1	0	1	1	1	0	1	9
6	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	12
7	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	7
8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
9	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	12
10	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	13

Keterangan: 1 = butir soal dijawab benar, 0 = butir soal dijawab salah.

Perhatikan Tabel 5.1. Untuk butir soal nomor 1, banyaknya yang menjawab benar butir itu adalah 6, sehingga B = 6. Banyaknya seluruh peserta tes adalah 10, sehingga N = 10. Dengan demikian untuk butir soal nomor 1 diperoleh $P_1 = \frac{6}{10} = 0.6$.

Berdasarkan Tabel 5.1 itu pula dapat diperoleh indeks tingkat kesulitan untuk butir lain sebagai berikut.

$$P_2 = \frac{6}{10} = 0{,}6,\ P_3 = \frac{9}{10} = 0{,}9, \ldots,\ P_{15} = \frac{8}{10} = 0{,}8$$

Misalnya diberikan batasan bahwa butir soal yang baik dari sisi tingkat kesulitan adalah $0{,}30 \leq P \leq 0{,}70$, maka butir yang tidak baik adalah butir soal nomor 3, 5, 7, 8, 9, 11, 12, dan 15.

Misalnya diberikan batasan bahwa butir soal yang baik dari sisi tingkat kesulitan adalah $0{,}20 \leq P \leq 0{,}80$, maka butir yang tidak baik adalah butir soal nomor 3, dan 9.

DAYA PEMBEDA (*DISCRIMINATION POWER*)

Suatu butir soal mempunyai daya pembeda baik jika kelompok siswa pandai menjawab benar butir soal lebih banyak daripada kelompok siswa tidak pandai. Dengan demikian, daya pembeda suatu butir soal dapat dipakai untuk membedakan siswa yang pandai dan tidak pandai. Sebagai tolok ukur pandai atau tidak pandai adalah skor total dari sekumpulan butir yang dianalisis.

Ada beberapa cara untuk mengukur daya pembeda, yaitu sebagai berikut.

Cara Pertama (Cara Klasik)

Dengan cara ini, peserta tes diurutkan dari skor total tertinggi sampai dengan skor total terrendah. Berdasarkan aturan tertentu, peserta tes dikelompokkan menjadi dua kelompok, yaitu kelompok atas (pandai) dan kelompok bawah (tidak pandai). Biasanya penentuan itu didasarkan atas mediannya, yang berarti separuh dari peserta tes adalah kelompok atas dan separuh dari peserta tes adalah kelompok bawah. Jika banyak datanya ganjil, maka data yang berada di tengah tidak diikutkan dalam analisis.

Jika peserta tesnya dalam jumlah besar, dapat digunakan aturan bahwa 27% (atau 30%) urutan teratas adalah kelompok atas dan 27% (atau 30%) urutan terbawah adalah kelompok bawah. Hal ini didasarkan pada pengalaman empirik bahwa 27% (atau 30%) kelompok atas dan 27% (atau 30%) kelompok bawah dapat mewakili separuh kelompok atas dan separuh kelompok bawah.

Indeks daya pembeda dirumuskan sebagai berikut.

$$D = \frac{B_a}{N_a} - \frac{B_b}{N_b} \qquad 5.2$$

dengan D adalah indeks daya pembeda butir soal. B_a adalah banyaknya peserta tes pada kelompok atas yang menjawab benar, N_a adalah banyaknya peserta tes pada kelompok atas, B_b adalah banyaknya peserta tes pada kelompok bawah yang menjawab benar, dan N_b adalah banyaknya peserta tes pada kelompok bawah.

Jika pembagian menjadi kelompok atas dan kelompok bawah didasarkan kepada median, maka banyaknya peserta tes pada kelompok atas sama dengan banyaknya peserta tes pada kelompok bawah. Jika pembagiannya didasarkan atas rerata, maka bisa jadi banyaknya peserta tes pada kelompok atas tidak sama dengan banyaknya peserta tes pada kelompok bawah.

Contoh 5.2

Suatu tes pilihan ganda terdiri dari 15 butir dikenakan kepada 10 siswa. Sebaran skor untuk masing-masing butir dan skor total peserta tes tampak pada Tabel 5. 1 di depan.

Untuk mencari indeks daya pembeda dengan cara pertama, peserta tes diurutkan dari skor total tertinggi ke terrendah. Kemudian, berdasarkan mediannya, peserta tes dikelompokkan menjadi kelompok atas dam kelompok bawah seperti pada Tabel 5.2.

Tabel 5.2. Skor 10 Siswa setelah Diurutkan

Nomor Urut Siswa	Nomor Butir Soal															Skor Total	Kelompok
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15		
8	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15	Atas
3	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	14	Atas
10	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	13	Atas
9	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	12	Atas
6	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	12	Atas
4	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	11	Bawah
1	1	1	1	0	1	0	0	1	1	1	0	1	1	0	1	10	Bawah
5	0	1	0	0	0	1	1	1	1	0	1	1	1	0	1	9	Bawah
7	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	7	Bawah
2	0	0	1	1	0	0	0	1	1	0	1	0	0	1	0	6	Bawah

Perhatikan Tabel 5.2. Untuk butir soal nomor 1, misalnya, indeks daya pembeda dapat dicari dengan cara berikut. Banyaknya siswa kelompok atas yang menjawab benar ada orang, sehingga $B_a = 5$ dan banyaknya siswa kelompok bawah yang menjawab benar ada 2, sehingga $B_b = 2$. Berdasarkan ini diperoleh:

$$D_1 = \frac{B_a}{N_a} - \frac{B_b}{N_b} = \frac{5}{5} - \frac{1}{5} = 0{,}8$$

Dengan cara yang sama, diperoleh:

$D_2 = 0$, $D_3 = 0{,}2$, $D_4 = 0.6$, $D_5 = 0{,}2$, $D_6 = 0{,}6$, $D_7 = 0.4$, $D_8 = 0$, $D_9 = -0{,}2$, $D_{10} = 0{,}2$, $D_{11} = 0$, $D_{12} = 0{,}4$, $D_{13} = 0.2$, $D_{14} = 0.6$, dan $D_{15} = 0.4$.

Perhatikan kembali indeks daya pembeda dirumuskan sebagai berikut.

$$D = \frac{B_a}{N_a} - \frac{B_b}{N_b}$$

Perhatikan bahwa $\frac{B_a}{N_a}$ merupakan tingkat kesulitan butir pada siswa-siswa kelompok atas, sedangkan $\frac{B_b}{N_b}$ merupakan tingkat kesulitan pada siswa-siswa kelompok bawah. Dengan demikian, indeks daya beda suatu butir dapat dicari dari formula berikut.

$$D = P_a - P_b \qquad 5.3$$

dengan P_a adalah tingkat kesulitan butir soal pada kelompok atas dan P_b adalah tingkat kesulitan butir soal pada kelompok bawah.

Dengan demikian, indeks daya pembeda suatu butir dapat dirumuskan seperti pada Persmaaan 5.3

Rentang Indeks Daya Pembeda

Perhatikan kembali Persamaan 5.2.

$$D = \frac{B_a}{N_a} - \frac{B_b}{N_b}$$

Jika $B_a = 0$ dan $B_b = N_b$ (yang berarti tidak ada peserta tes pada kelompok atas yang menjawab benar dan semua peserta tes pada kelompok bawah menjawab benar), maka $D = -1$. Sebaliknya, jika $B_a = N_a$ dan

$B_b = 0$ (yang berarti semua peserta tes pada kelompok atas menjawab benar dan semua peserta tes pada kelompok bawah tidak ada yang menjawab benar), maka D = 1. Dengan demikian, rentang indeks daya pembeda adalah $-1 \leq D \leq 1$.

Cara Kedua (dengan Koefisien Korelasi Biserial Titik)

Perhatikan kembali Tabel 5.2. Pada Tabel 5.3 dicuplikkan skor butir yang mempunyai daya pembeda positif (yaitu butir soal nomor 1 dan 4) dan yang mempunyai daya pembeda negatif (yaitu butir soal nomor 9), dan skor totalnya serta pembagian kelompok atas-bawahnya.

Tabel 5.3. Sebaran Skor Butir 1, 4, 9 dan Sebaran Skor Total

Nomor Urut Siswa	1	4	9	Skor Total	Kelompok
8	1	1	1	15	**Atas**
3	1	1	1	14	**Atas**
10	1	1	1	13	**Atas**
9	1	1	1	12	**Atas**
6	1	1	0	12	**Atas**
4	0	1	1	11	Bawah
1	1	0	1	10	Bawah
5	0	0	1	9	Bawah
7	0	0	1	7	Bawah
2	0	1	1	6	Bawah
..	D=0.8	D=0,6	D=-0.2		

Perhatikan sebaran skor butir soal nomor 1 dan nomor 4 dan skor totalnya. Terdapat kecenderungan bahwa kelompok atas cenderung menjawab benar dan kelompok bawah cenderung menjawab salah. Ini berarti pada butir yang indeks daya pembedanya positif, terdapat korelasi positif antara skor butir dengan skor totalnya. Di sisi lain, perhatikan sebaran skor butir nomor 9 di mana semua peserta tes kelompok bawah menjawab benar dan tidak semua peserta tes kelompok atas menjawab benar. Ini berarti terdapat korelasi negatif antara skor butir nomor 9 dengan skor totalnya. Indeks daya pembeda butir nomor 9 negatif. Dengan demikian, ada cara lain untuk

mencari indeks daya pembeda, yaitu dengan mencari koefisien korelasi antara skor butir dan skor total[2].

Pada cara kedua ini, indeks daya pembeda suatu butir dicari dengan mencari koefisien korelasi antara skor butir tersebut dengan skor total peserta tes. Dengan demikian, indeks daya pembeda dirumuskan sebagai berikut.

$$D = r_{pbis} = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}} \quad 5.3$$

dengan X adalah skor butir dan Y adalah skor total

Cara kedua ini disebut cara dengan menggunakan koefisien korelasi biserial titik (*point biserial correlation*).

Contoh 5.3

Untuk mencari daya pembeda pada butir nomor 1 pada contoh di atas dengan koefisien korelasi biserial titik dapat dicari sebagai berikut.

Tabel 5.4. Tabel Kerja untuk Menghitung Daya Pembeda Butir Nomor 1

											Total
Skor Butir ke-1 (X)	1	0	1	0	0	1	0	1	1	1	6
Skor Total (Y)	10	6	14	11	9	12	7	15	12	13	109
X^2	1	0	1	0	0	1	0	1	1	1	6
Y^2	100	36	196	121	81	144	49	225	144	169	1265
XY	10	0	14	0	0	12	0	15	12	13	76

$$D_1 = r_{pbis} = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}}$$

$$= \frac{(10)(76) - (6)(109)}{\sqrt{((10)(6) - 6^2)((10)(1265) - 109^2)}} = 0,78$$

Jadi, dihitung dengan koefisien korelasi biserial titik, maka indeks daya pembeda butir soal nomor 1 adalah $D_1 = 0.78$.

Dengan cara yang sama, diperoleh:

$D_2 = 0.27$, $D_3 = 0.23$, $D_4 = 0.53$, $D_5 = 0.61$,
$D_6 = 0.76$, $D_7 = 0,52$, $D_8 = 0.25$, $D_9 = 0,135$,
$D_{10} = 0,54$, $D_{11} = 0,02$, $D_{12} = 0,79$, $D_{13} = 0,29$,
$D_{14} = 0,40$, dan $D_{15} = 0,43$.

Cara Ketiga (dengan Koefisien Korelasi Biserial Titik)

Rumus pada cara kedua dapat disederhanakan dalam rumus berikut ini.

$$D = r_{pbis} = \left(\frac{\overline{Y}_1 - \overline{Y}}{\sigma Y}\right)\sqrt{\frac{p_X}{(1-p_X)}} \qquad 5.4$$

dengan X adalah skor butir, Y adalah skor total, $\overline{Y}_1$ adalah rerata skor Y dengan X = 1, $\overline{Y}$ adalah rerata untuk skor total untuk Y, σ_Y adalah deviasi baku dari skor total (dianggap populasi) dengan $\sigma_Y^2 = \frac{\Sigma Y^2}{N} - \left(\frac{\sum Y}{N}\right)^2$, N adalah banyaknya siswa, dan p_X adalah proporsi peserta tes dengan X = 1.

Contoh 5.4

Dengan rumus koefisien korelasi biserial titik yang kedua, indeks daya pembeda butir soal nomor 1 dihitung sebagai berikut.

$$\Sigma Y = 10 + 6 + 14 + \ldots + 12 + 13 = 109$$

$$\sum Y^2 = 10^2 + 6^2 + 14^2 + \ldots + 12^2 + 13^2 = 1265$$

$$\overline{Y}_1 = \frac{10+14+12+15+12+13}{6} = 12,667;$$

$$\overline{Y} = \frac{10+6+14+\ldots+12+13}{10} = 10,900;$$

$$\sigma_Y = \sqrt{\frac{1265}{10} - \left(\frac{109}{10}\right)^2} = 2,733;$$

$$p_x = \frac{6}{10} = 0,6$$

$$D_1 = \left(\frac{\overline{Y}_1 - \overline{Y}}{\sigma Y}\right)\sqrt{\frac{p_X}{(1-p_X)}} = \left(\frac{12.667-10.900}{2.733}\right)\sqrt{\frac{0,6}{(1-0.6)}} = 0,78$$

Cara Keempat (dengan Koefisien Korelasi Biserial)

Rumus pada cara keempat ini adalah sebagai berikut.

$$D = r_{bis} = \left(\frac{\overline{Y}_1 - \overline{Y}}{\sigma Y}\right)\frac{p_X}{f(z)}$$

dengan X adalah skor butir, Y adalah skor total, $\overline{Y}_1$ adalah rerata skor Y dengan X = 1, $\overline{Y}$ adalah rerata untuk skor total untuk Y, σ_Y adalah deviasi baku dari skor total (dianggap populasi) dengan $\sigma_Y^2 = \frac{\Sigma Y^2}{N} - \left(\frac{\Sigma Y}{N}\right)^2$. N adalah banyaknya siswa, p_X adalah proporsi peserta tes dengan X = 1, z adalah nilai pada distribusi normal baku demikian hingga luas di bawah kurva normal baku dan di sebelah kanan z adalah p_x, dan $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$.

Contoh 5.5

Dengan rumus koefisien korelasi biserial, indeks daya pembeda butir soal nomor 1 dihitung sebagai berikut.

$\overline{Y}_1 = 12,667$; $\overline{Y} = 10,900$; $\sigma_Y = 2,733$; $p_x = 0,6$;

$z = -0,25$ (diperoleh dari tabel distribusi normal baku);

$$f(-0,25) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} = \left(\frac{1}{\sqrt{(2)(3,143)}}\right)\left(2,714^{-\frac{1}{2}(-0,25)^2}\right) = 0,39,$$

sehingga:

$$D_1 = r_{bis} = \left(\frac{\overline{Y}_1 - \overline{Y}}{\sigma Y}\right)\frac{p_X}{f(z)}$$

$$= \left(\frac{12,667-10,900}{2,733}\right)\left(\frac{0,6}{0,39}\right) = (0.647)(1,600) = 0,99$$

Di antara keempat cara di atas, dulu ketika alat kalkulasi belum banyak tersedia, orang menggunakan cara pertama. Bahkan, pengambilan 27% urutan teratas untuk kelompok atas dan 27% urutan terbawah untuk kelompok bawah adalah untuk menyederhanakan perhitungan. Namun

sekarang ini, dengan telah tersedianya paket pengolah data, misalnya MS Exel, atau paket program statistik, misalnya SPSS, maka cara kedua, yaitu dengan koefisien korelasi biserial titik yang dianjurkan. Cara ini dianjurkan karena di samping dapat untuk tes pilihan ganda, dapat juga digunakan untuk tes uraian.

Indeks Daya Pembeda yang Diperbolehkan

Pengembang tes biasanya ingin memperoleh daya pembeda yang positif dalam arti kelompok pandai menjawab benar butir soal lebih banyak daripada kelompok tidak pandai. Jika dianalisis dari definisi daya pembeda, maka semakin mendekati 1, semakin baik indeks daya pembeda untuk butir tersebut. Namun demikian, mengupayakan indeks daya pembeda yang sangat tinggi, cukup sukar.

Oleh karena itu, biasanya, suatu butir soal dikatakan mempunyai daya beda yang baik apabila indeks daya bedanya sama atau lebih dari 0,30. (jadi, $D \geq 0,30$).

BERFUNGSINYA PENGECOH

Pengecoh yang baik harus dipilih oleh peserta tes. Untuk menentukan apakah pengecoh berfungsi atau tidak, biasanya, diambil nilai ambang 5%. Artinya, salah satu syarat agar pengecoh dikatakan berfungsi baik adalah jika pengecoh tersebut dipilih oleh paling sedikit 5% peserta tes.

Agar dapat mengecoh peserta tes, maka pengecoh-pengecoh yang ada pada suatu butir soal harus sama kuat daya tariknya. Suatu pengecoh yang sangat berbeda dengan pengecoh lainnya tentu saja tidak dianjurkan. Perhatikan contoh butir soal berikut.

Contoh 5.6

Perhatikan butir soal berikut.

Raja pertama Majapahit adalah ... .
a. Tunggul Ametung
b. Brawijaya
c. Hayam Wuruk
d. Ken Arok
e. Superman

Pengecoh d tidak mempunyai daya tarik sama kuat dengan pengecoh lainnya, karena pengecoh d bukan nama raja, dan semua orang tahu bahwa Superman bukanlah nama raja, tidak seperti pengecoh yang lainnya. Oleh karena itu, pengecoh d harus diganti.

Pada mata pelajaran tertentu, seperti misalnya matematika, pengecoh disusun berdasarkan atas kesalahan yang mungkin dilakukan oleh peserta tes.

Contoh 5.7

Misalnya *stem*nya adalah "Himpunan penyelesaian persamaan kuadrat $4x^2 - 1 = 0$ adalah ... ."

Kunci jawaban dari butir soal tersebut dapat dicari dari pengerjaan berikut.

$$4x^2 - 1 = 0$$
$$\Leftrightarrow (2x + 1)(2x - 1) = 0$$
$$\Leftrightarrow 2x + 1 = 0 \text{ atau } 2x - 1 = 0$$
$$\Leftrightarrow x = -\frac{1}{2} \text{ atau } x = \frac{1}{2}$$
$$HP = \{ \frac{1}{2}, -\frac{1}{2} \}$$

Untuk menentukan pengecoh dari *stem* tersebut, dipikirkan kesalahan yang mungkin dilakukan oleh siswa. Diduga ada siswa yang mengerjakan soal tersebut dengan cara-cara seperti berikut.

$$4x^2 - 1 = 0$$
$$\Leftrightarrow (4x + 1)(4x - 1) = 0$$
$$\Leftrightarrow 4x + 1 = 0 \text{ atau } 4x - 1 = 0$$
$$\Leftrightarrow x = -\frac{1}{4} \text{ atau } x = \frac{1}{4}$$
$$HP = \{ \frac{1}{4}, -\frac{1}{4} \}$$

$$4x^2 - 1 = 0$$
$$\Leftrightarrow 4x^2 = 1$$
$$\Leftrightarrow x^2 = \frac{1}{4}$$
$$\Leftrightarrow x = \frac{1}{2}$$
$$HP = \{ \frac{1}{2} \}$$

$$4x^2 - 1 = 0$$
$$\Leftrightarrow 4x^2 = 1$$
$$\Leftrightarrow x^2 = \frac{1}{4}$$
$$\Leftrightarrow x = \frac{1}{4}$$
$$HP = \{\frac{1}{4}\}$$

$$4x^2 - 1 = 0$$
$$\Leftrightarrow 3x^2 = 0$$
$$\Leftrightarrow x^2 = 0$$
$$\Leftrightarrow x = 0$$
$$HP = \{0\}$$

Dengan pemikiran seperti itu, maka butir soal tersebut dapat disusun seperti berikut.

Himpunan penyelesaian persamaan kuadrat $4x^2 - 1 = 0$ adalah ....
a. $\{0\}$
b. $\{\frac{1}{4}\}$
c. $\{\frac{1}{2}\}$
d. $\{\frac{1}{4}, -\frac{1}{4}\}$
e. $\{\frac{1}{2}, -\frac{1}{2}\}$

Kecuali dipilih oleh paling sedikit 5% dari seluruh peserta tes, pengecoh yang baik harus lebih mengecoh kelompok bawah daripada kelompok atas. Artinya, supaya pengecoh berfungsi, peserta tes kelompok bawah yang memilih pengecoh tersebut harus lebih banyak daripada peserta tes kelompok atas. Perhatikan contoh berikut.

Contoh 5.8

Berikut ini terdapat sebaran jawaban sekelompok peserta tes untuk butir soal tertentu.

Tabel 5.5. Sebaran Jawaban Peserta Tes untuk Butir Soal Tertentu

Kelompok	Pilihan Jawaban				
	A	B	C	D	E
Kelompok Atas	1	5	42	4	0
Kelompok Bawah	9	5	26	3	9

Keterangan: **kunci jawaban C**

Butir soal tersebut mempunyai indeks tingkat kesulitan $P = \frac{68}{104} = 0{,}65$ dan $D = \frac{42}{52} - \frac{26}{52} = 0{,}31$, yang berarti merupakan butir soal yang cukup baik untuk mengambil data prestasi belajar pada suatu penelitian, sebab $0{,}30 \leq P \leq 0{,}70$ dan $D \geq 0.30$. Namun demikian, pengecoh B dan pengecoh D tidak berfungsi dengan baik, sebab kelompok bawah tidak lebih banyak yang memilih pengecoh-pengecoh tersebut dibandingkan dengan kelompok atas.

Perhatikan baik-baik bahwa semakin baik pegecoh berfungsi, butir soal tersebut semakin mempunyai daya pembeda yang baik, namun demikian, indeks tingkat kesulitannya cenderung menurun (berarti butir soal semakin sulit).

PAKET PROGRAM UNTUK ANALISIS BUTIR

Dewasa ini banyak paket program komputer yang ditawarkan untuk melakukan analisis butir. Salah satu di antaranya adalah paket program komputer yang diberi nama ITEMAN, singkatan dari *item analysis*. Paket program ITEMAN dibuat oleh *Assessment Systems Corporation* di Amerika Serikat.

Untuk ITEMAN Versi 3.0 (yang dipunyai oleh penulis) masih menggunakan sistem operasi DOS. Pada versi tersebut, file yang harus dieksekusi diberi nama **iteman.exe**. File input data yang akan dianalisis diketik dalam file ASCII atau DOS Text format, dengan ekstension **.dat**. Untuk memperoleh file tersebut dapat digunakan perintah **edit** pada DOS command atau dapat digunakan **Notepad**.

Ketentuan untuk menulis file input adalah sebagai berikut.

1. Baris pertama berisi kode-kode sebagai berikut.

Kolom	Keterangan	Contoh
1 – 3	Banyaknya butir yang dianalisis	020
4	Kosong/spasi	
5	Untuk jawaban *omit*/kosong	O
6	Kosong/spasi	
7	Untuk butir soal yang tidak (belum sempat) dikerjakan	N
8	Kosong/spasi	
9 – 10	Banyaknya kolom yang diperlukan untuk identitas peserta tes	4

2. Baris kedua berisi kunci jawaban
3. Baris ketiga berisi banyaknya alternatif jawaban
4. Baris keempat: berisi kode: "Y" berarti butir dianalisis. "N" butir tidak dianalisis.

Ada dua *file* output pada program ITEMAN. *File* pertama berisi hasil analisis butir, *file* kedua berisi skor peserta tes.

Contoh 5.9

Misalnya terdapat perangkat tes yang terdiri dari 20 butir soal yang diberikan kepada 10 peserta tes. Misalnya sebaran jawaban siswa adalah sebagai berikut.

Tabel 5.6. Sebaran Jawaban 10 Peserta Tes untuk 20 Butir Soal

No Siswa	Nomor Butir Soal																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	A	C	B	E	E	D	C	E	B	D	C	C	D	C	A	A	B	A	A	C
2	A	A	B	D	B	B	C	E	E	D	C	B	D	C	A	A	B	A	C	D
3	C	A	B	D	B	B	C	E	E	B	C	E	D	C	A	A	B	A	B	D
4	C	A	B	E	B	B	C	E	E	B	C	E	D	C	A	A	O	A	D	O
5	C	B	A	E	B	C	D	E	E	A	B	C	D	A	A	A	B	A	C	D
6	D	A	B	D	B	E	D	E	C	B	D	C	D	C	A	A	C	B	E	E
7	C	A	C	D	B	B	B	E	E	B	C	E	D	C	A	A	B	A	D	C
8	C	A	B	D	B	B	B	E	E	D	C	E	D	C	B	C	C	B	A	A
9	B	D	C	D	C	C	B	C	D	A	D	B	B	D	C	C	C	A	C	D
10	A	B	C	B	E	D	E	E	A	A	C	D	A	B	A	A	B	A	C	C

Misalnya, file *input* datanya diberi nama **data.dat** dan ditulis sebagai berikut[3].

[3] Perhatikan bahwa *file* ini harus ditulis dalam sistem ASCI. misalnya dengan menggunakan Notepad atau perintah Edit pada sistem DOS.

```
020 O N   4
CCBDBDEEEBCADCAABACE
55555555555555555555
YYYYYYYYYYYYYYYYYYYY
001  ACBEEDCEBDCCDCAABAAC
002  AABDBBCEEDCBDCAABACD
003  CABDBBCEEBCEDCAABABD
004  CABEBBCEEBCEDCAAOADC
005  CBAEBCDEEABCDAAABACD
006  DABDBEDECBDCDCAACBEE
007  CACDBBBEEBCEDCAABADC
008  CABDBBBEEDCEDCBCCBAA
009  BDCDCCBCDADBBDCCCACD
010  ABCBEDEEAACDABAABACC
```

Setelah program dieksekusi, maka hasilnya dapat dilihat pada *file* output. Misalnya file tersebut disebut **output**, maka pada file tersebut dapat dilihat karakteristik masing-masing butir. Untuk butir soal nomor 1, misalnya, diperoleh keluaran berikut.

```
                 Item Statistics                  Alternative Statistics
           -----------------------------   -------------------------------------
Seq  Scale   Prop.                Point             Prop.                Point
No.  -Item   Correct    Biser.    Biser.   Alt.     Endorsing   Biser.   Biser   Key
---  -----   -------    -----     ------   ----     ---------   ------   ------  --
  1   0-1     0.500     0.506     0.404     A        0.300       0.174    0.132
                                            B        0.100      -1.000   -0.830
                                            C        0.500       0.506    0.404   *
                                            D        0.100      -0.077   -0.045
                                            E        0.000      -9.000   -9.000
                                          Other      0.000      -9.000   -9.000
```

Berdasarkan keluaran tersebut, diperoleh hal-hal berikut:

1. Indeks tingkat kesulitan P = 0,500
2. Indeks daya pembeda D = 0.506 (jika menggunakan korelasi biserial) atau D = 0,404 (jika menggunakan korelasi biserial titik).
3. Untuk pengecoh A, banyaknya peserta tes yang memilihnya ada 30% dan diperoleh indeks daya pembeda D = 0,132. Karena indeks daya pembedanya positif, berarti pengecoh A dipilih lebih banyak kelompok atas daripada kelompok bawah. Ini berarti pengecoh A tidak berfungsi.

4. Untuk pengecoh B, banyaknya peserta tes yang memilihnya ada 10% dan diperoleh indeks daya pembeda D = -0,830. Karena indeks daya pembedanya negatif, berarti pengecoh B dipilih lebih banyak kelompok bawah daripada kelompok atas. Ini berarti pengecoh B berfungsi dengan baik.
5. Dengan pemikiran yang sama seperti pada pengecoh B, maka pengecoh D juga berfungsi baik.
6. Pengecoh E tidak dipilih lebih dari 5% peserta (karena banyak pemilihnya 0%), maka pengecoh E tidak berfungsi.

Berikut ini adalah keluaran untuk butir soal nomor 20.

Item Statistics					Alternative Statistics				
Seq No.	Scale -Item	Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser	Key
20	0-20	0.100	-0.077	-0.045	A	0.100	-0.268	-0.157	
					B	0.000	-9.000	-9.000	
	CHECK THE KEY				C	0.300	0.174	0.132	?
E was specified, C works better					D	0.400	-0.139	-0.110	
					E	0.100	-0.077	-0.045	*
					Other	0.100	0.307	0.179	

Berdasarkan keluaran tersebut, diperoleh hal-hal berikut:

1. Indeks tingkat kesulitan P = 0,100, yang berarti butir soal terlalu sulit.
2. Indeks daya pembeda D = −0,077 (jika menggunakan korelasi biserial) atau D = −0,045 (jika menggunakan korelasi biserial titik). Ini berarti bahwa butir soal nomor 20 bukan butir soal yang baik, karena daya pembedanya negatif.
3. Untuk pengecoh A, banyaknya peserta tes yang memilihnya ada 10% dan diperoleh indeks daya pembeda D = −0,157. Berarti, pengecoh A dipilih lebih banyak kelompok bawah daripada kelompok atas. Ini berarti pengecoh A berfungsi.
4. Untuk pengecoh B, tidak ada peserta tes yang memilihnya. Berarti, bukan pengecoh yang baik.
5. Pengecoh C adalah pengecoh yang tidak baik karena daya pembedanya positif, walau dipilih oleh 30% peserta tes.

6. Pengecoh E adalah pengecoh yang baik karena dipilih oleh 10% dan daya pembedanya negatif.
7. Terdapat 10% peserta yang tidak mengerjakan (atau tidak *option* A, B, C, D, maupun E). Lihat alternatif Other.
8. Perhatikan bahwa ITEMAN memberikan masukan agar pengembang tes meninjau kembali kunci jawaban. Kunci jawaban yang disebut oleh pengembang adalah E, namun ITEMAN menyarankan apakah kunci jawabannya bukan C, karena pengecoh C mempunyai daya pembeda positif yang paling besar di antara daya pembeda yang lain.

Pada akhir *file* output untuk hasil analisis butir juga dimunculkan ringkasan analisis seperti berikut.

```
N of Items              20
N of Examinees          10
Mean                10.400
Variance             8.840
Std. Dev.            2.973
Skew                -1.235
Kurtosis             1.188
Minimum              3.000
Maximum             14.000
Median              10.000
Alpha                0.638
SEM                  1.788
Mean P               0.520
Mean Item-Tot.       0.362
Mean Biserial        0.468
```

Dari keluaran itu dapat dilihat, misalnya, koefisien reliabilitas tes, yang dihitung dengan teknik alpha adalah sebesar 0,638.

ANALISIS BUTIR UNTUK SOAL URAIAN

Berbeda dengan analisis butir untuk soal pilihan ganda, tidak banyak buku yang membicarakan analisis butir untuk soal bentuk uraian. Pada buku ini, analisis butir untuk soal bentuk uraian dikembangkan dari analisis butir untuk soal bentuk pilihan ganda.

TINGKAT KESULITAN

Indeks tingkat kesulitan untuk tes uraian dirumuskan sebagai berikut.

$$P = \frac{\bar{S}}{S_{maks}} \qquad 5.5$$

dengan P adalah indeks tingkat kesulitan, $\bar{S}$ adalah rerata untuk skor butir, dan S_{maks} adalah skor maksimum untuk butir tersebut.

Contoh 5.10

Misalnya terdapat 5 butir soal bentuk uraian yang dikenakan pada 10 orang siswa. Setiap butir diskor dengan skala 10 (skor minimal 1 dan skor maksimal 10). Sebaran skor mereka adalah sebagai berikut.

Tabel 5.7. Sebaran Skor untuk 10 Peserta Tes pada 5 Butir Uraian

No Butir	Nama Siswa									
	Aa	Bb	Cc	Dd	Ee	Ff	Gg	Hh	Ii	Jj
1	6	9	7	9	7	4	7	6	5	5
2	7	8	7	9	7	5	8	7	7	3
3	6	9	6	9	7	5	7	6	8	4
4	5	7	8	10	8	4	8	6	6	4
5	8	9	7	9	7	6	7	8	7	6
Skor Total	32	42	35	46	36	24	37	33	33	22

Dalam kasus ini, skor maksimal untuk masing-masing butir soal adal 10, sehingga indeks tingkat kesulitan untuk butir soal nomor 1 dicari sebagai berikut.

$$P_1 = \frac{\bar{S}}{S_{maks}} = \frac{\frac{65}{10}}{10} = \frac{6{,}5}{10} = 0{,}65$$

Dengan cara yang sama, diperoleh: $P_2 = 0{,}68$; $P_3 = 0{,}67$; $P_4 = 0{,}66$; dan $P_5 = 0{,}74$.

DAYA PEMBEDA

Indeks daya pembeda dicari dengan mencari koefisien korelasi antara skor butir dengan skor total sebagai berikut.

$$D = r_{pbis} = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}} \quad 5.6$$

dengan X adalah skor butir dan Y adalah skor total

Contoh 5.11
Untuk menghitung indeks daya pembeda untuk butir soal pertama, dicari koefisien korelasi antara skor butir pertama dengan skor total, sebagai berikut.

Tabel 5.8. Tabel Kerja untuk Mencari Daya Pembeda Soal Nomor 1

Nomor Urut Siswa	1	2	3	4	5	6	7	8	9	10	Total
Skor Butir ke-1 (X)	6	9	7	9	7	4	7	6	5	5	65
Skor Total Siswa (Y)	32	42	35	46	36	24	37	33	33	22	340
XY	192	378	245	414	252	96	259	198	165	110	2309
X^2	36	81	49	81	49	16	49	36	25	25	447
Y^2	1024	1764	1225	2116	1296	576	1369	1089	1089	484	12032

Indeks daya pembeda untuk butir soal nomor 1 dicari sebagai berikut.

$$D_1 = r_{pbis} = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}}$$

$$= \frac{(10)(2309) - (65)(340)}{\sqrt{\left((10)(447) - 65^2\right)\left((10)(12032) - 340^2\right)}} = 0,92$$

Dengan cara yang sama, diperoleh $D_2 = 0,94$; $D_3 = 0,91$; $D_4 = 0,90$; dan $D_5 = 0,84$.

BAHAN DISKUSI

1. a. Apa yang disebut dengan indeks kesulitan butir soal? Jelaskan!
   b. Apakah semakin tinggi indeks kesulitan butis soal, butir soal tersebut semakin sulit? Mengapa?
   b. Apa yang disebut dengan indeks daya beda butir soal? Jelaskan!
   c. Misalnya suatu butir soal mempunyai indeks daya pembeda D = 1. Amir menjawab benar butir soal tersebut dan Siti menjawab salah butir soal tersebut. Apakah dapat dipastikan bahwa Amir termasuk anak pandai dan Siti termasuk anak yang tidak pandai? Mengapa?

d. Misalnya suatu butir soal mempunyai indeks daya pembeda D = 0. Parti menjawab benar butir soal tersebut dan Wanti menjawab salah butir soal tersebut. Apakah dapat dipastikan kalah Parti termasuk anak pandai dan Wanti termasuk anak yang tidak pandai? Mengapa?
e. Apa artinya jika daya beda suatu butir soal negatif? Jelaskan!
f. Apa artinya jika tingkat kesukaran butir soal negatif? Jelaskan!
g. Suatu butir soal mempunyai D = 0.45 dan P = 0.25. Apakah butir tersebut merupakan butir yang baik? Mengapa?

2. Menurut Anda, manakah yang lebih menguntungkan siswa:
   a. butir soal yang indeks tingkat kesulitannya rendah
   b. butir soal yang indeks tingkat kesulitannya tinggi
   c. butir soal yang indeks daya pembedanya di sekitar nol
   d. butir soal yang indeks daya pembedanya mendekati satu.

3. Berikut ini adalah sebaran skor 8 siswa pada 15 butir soal pilihan ganda.

Tabel 5.11. Sebaran Skor 8 Siswa dalam Menjawab 15 Butir Soal

Nomor Urut Siswa	Nomor Butir Soal															Skor Total
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
1	1	1	1	0	1	0	0	1	1	1	0	1	1	0	1	10
2	0	0	1	1	0	0	0	1	1	0	1	0	0	1	0	6
3	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	14
4	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	11
5	0	1	0	0	0	1	1	1	1	0	1	1	1	0	1	9
6	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	12
7	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	7
8	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	13

Keterangan: 1 = butir soal dijawab benar. 0 = butir soal dijawab salah.

a. Hitunglah indeks tingkat kesulitan dan indeks daya pembeda (dengan rumus pertama) masing-masing butir soal pada Tabel 5.11.
b. Dengan mengacu kepada ketentuan bahwa butir yang baik adalah butir yang $0{,}30 \leq P \leq 0{,}70$ dan $D \geq 0{,}30$, adakah butir yang baik pada data Tabel 5.11? Yang mana?

4. Diketahui data pada Tabel 5.11. Carilah daya beda masing-masing butir dengan rumus korelasi biserial titik. Apakah nilai sama persis dengan

nilai daya pembeda yang dicari dengan menggunakan rumus pertama? Mengapa?

5. Berapa rentang indeks daya pembeda jika digunakan rumus koefisien korelasi biserial titik? Jelaskan pendapat Anda!

6. Berikut ini terdapat sebaran jawaban sekelompok peserta tes untuk butir soal tertentu.

Kelompok	Pilihan Jawaban				
	A	B	C	D	E
Kelompok Atas	0	5	42	2	1
Kelompok Bawah	10	5	26	5	4

Keterangan: **kunci jawaban C**

Jika seluruh peserta tes menjawab butir soal tersebut, jawablah pertanyaan berikut.

a. Berapa tingkat kesulitan butir soal tersebut?
b. Berapa daya beda butir soal tersebut?
c. Apakah butir soal tersebut merupakan butir soal yang baik, jika dilihat tingkat kesulitan dan daya bedanya? Mengapa?
d. Mana saja pengecoh yang berfungsi? Mengapa?

7. Berikut ini terdapat sebaran jawaban sekelompok peserta tes untuk butir soal tertentu.

Kelompok	Pilihan Jawaban				
	A	B	C	D	E
Kelompok Atas	6	4	37	3	0
Kelompok Bawah	14	4	30	2	0

Banyaknya seluruh peserta tes adalah 100 orang. **Kunci jawabannya adalah C.**

a. Berapa tingkat kesulitan butir soal tersebut?
b. Berapa daya beda butir soal tersebut?
c. Apakah butir soal tersebut merupakan butir soal yang baik, jika dilihat tingkat kesulitan dan daya bedanya? Mengapa?
d. Mana saja pengecoh yang berfungsi? Mengapa?

8. Berikut ini adalah sebaran dari 10 peserta tes pada 20 butir soal pilihan ganda.

No Siswa	Nomor Butir Soal																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	A	C	B	E	E	D	C	E	B	D	C	C	D	C	A	A	B	A	A	C
2	A	A	B	D	B	B	C	E	E	D	C	B	D	C	A	A	B	A	C	D
3	E	A	B	D	B	D	C	E	E	B	C	E	D	C	A	A	B	A	B	D
4	C	A	B	E	B	B	C	E	E	B	C	E	D	C	A	A	B	A	D	D
5	D	B	A	E	B	C	D	E	E	A	B	C	D	A	A	A	B	A	C	D
6	D	A	B	D	B	E	D	E	C	B	D	C	D	C	A	A	C	B	E	E
7	C	A	C	D	B	B	B	E	E	B	C	E	D	C	D	A	C	A	A	D
8	B	A	B	D	B	B	B	E	E	D	C	E	D	C	B	C	C	B	A	A
9	B	D	C	D	C	C	B	C	D	A	D	B	B	D	C	C	C	A	C	D
10	A	B	C	B	E	D	E	E	A	A	C	D	A	B	A	A	B	A	C	C
KJ	A	A	B	D	B	B	C	E	E	A	C	B	D	C	A	A	C	A	C	D

Keterangan: KJ = kunci jawaban

Dengan menggunakan ITEMAN, lakukan analisis butir pada data tersebut dan sebutkan mana-mana butir yang baik, dan mana-mana butir yang tidak baik.

9. Berikut ini adalah hasil dari pengolahan dengan ITEMAN

```
                 Item Statistics              Alternative Statistics
            -----------------------   -------------------------------
Seq. Scale  Prop.          Point          Prop.             Point
No.  -Item Correct Biser. Biser. Alt.Endorsing  Biser.  Biser. Key
---- ----- ------- ------ ------ ---- --------- -----   ------ ---
 5   0-5   0.450   0.736  0.586   A   0.200     -0.644  -0.451
                                  B   0.450      0.736   0.586  *
                                  C   0.100     -0.[illegible] -0.394
                                  D   0.100     -0.032  -0.019
                                  E   0.150      0.054   0.035
                              Other   0.000     -9.000  -9.000
```

a. Berapa tingkat kesulitan butir soal tersebut?
b. Berapa indeks daya beda butir soal tersebut?
c. Apakah butir soal tersebut merupakan butir soal yang baik, jika dilihat tingkat kesulitan dan daya bedanya? Mengapa?
d. Adakah pengecoh yang tidak berfungsi baik? Jelaskan mengapa Anda mengatakan bahwa pengecoh itu merupakan pengecoh yang tidak berfungsi baik, jika ada!

10. Berikut ini adalah hasil keluaran ITEMAN untuk soal nomor 19 dan 20.

		Item Statistics			Alternative Statistics				
Seq. No.	Scale -Item	Prop. Correct	Biser.	Point Biser.	Alt.	Prop. Endorsing	Biser.	Point Biser.	Key
19	0-19	0.600	0.345	0.272	A	0.133	-0.465	-0.294	
					B	0.067	0.086	0.045	
					C	0.600	0.345	0.272	*
					D	0.133	-0.052	-0.033	
					E	0.067	-0.258	-0.134	
					Other	0.000	-9.000	-9.000	
20	0-20	0.333	0.978	[illegible]	A	0.133	-0.413	-0.261	
					B	0.067	-0.430	-0.223	
					C	0.200	-0.239	[illegible]	
					D	0.267	-0.440	[illegible]	
					E	0.333	0.978	[illegible]	*
					Other	[illegible]	[illegible]	[illegible]	

a. Apakah butir nomor 19 memenuhi persyaratan sebagai butir yang baik dilihat dari daya beda dan tingkat kesukarannya? Mengapa?

b. Apakah butir nomor 20 memenuhi persyaratan sebagai butir yang baik dilihat dari daya beda dan tingkat kesukarannya? Menapa?

c. Pada butir nomor 19, adakah pengecoh yang tidak berfungsi? Yang mana? Mengapa?

d. Pada butir nomor 19, adakah pengecoh yang tidak berfungsi? Yang mana? Mengapa?

****

BAB VI
NON TES

PENDAHULUAN

Perhatikan kembali pengertian tes yang disampaikan pada Bab III. Tes didefinisikan sebagai seperangkat pertanyaan atau tugas yang direncanakan untuk memperoleh informasi tentang trait atau atribut pendidikan atau atribut psikologik tertentu yang setiap butir pertanyaan atau tugas tersebut mempunyai jawaban atau ketentuan yang dianggap benar. Respons peserta pada satu tes harus dapat dikategorikan sebagai respons yang benar atau respons yang salah. Jika ada pertanyaan atau tugas yang harus dikerjakan oleh seseorang, tetapi tidak ada jawaban atau cara mengerjakan yang benar atau salah, maka pertanyaan atau tugas tersebut bukanlah suatu tes dan disebut dengan non tes.

Pada non tes, tidak ada jawaban benar atau jawaban salah, tetapi dari respons peserta pada jawaban non tes dapat dilihat arah kecenderungannya. Itu berarti bahwa informasi mengenai hasil belajar tidak hanya dapat diperoleh melalui tes, tetapi dapat juga diperoleh melalui alat pengukuran yang disebut non-tes, seperti *rating scale* (skala lajuan), dan *attitude scale* (skala sikap).

Alat ukur untuk memperoleh informasi hasil belajar yang diungkap melalui non-tes terutama digunakan untuk mengetahui apa yang dilakukan siswa daripada apa yang diketahui atau dipahaminya. Alat ukur non-tes berhubungan dengan penampilan yang dapat diamati daripada pengetahuan dan proses mental lainnya yang tidak dapat diamati dengan indera manusia. Namun demikian, alat ukur non-tes ini merupakan satu kesatuan dengan alat ukur tes untuk memperoleh informasi hasil belajar yang lebih menyeluruh.

SKALA LAJUAN (*RATING SCALE*)

Skala lajuan adalah alat ukur non-tes yang menggunakan suatu prosedur terstruktur untuk memperoleh informasi mengenai sesuatu yang diobservasi yang menyatakan posisi sesuatu dalam hubungannya dengan sesuatu yang lain. Biasanya skala lajuan terdiri dari: (1) pernyataan tentang karakteristik atau kualitas sesuatu yang diukur dan (2) cara menilai yang menunjukkan peringkat atau karakter atau kualitas yang dimiliki oleh sesuatu tersebut.

Ada beberapa tipe skala lajuan, di antaranya: (1) *numerical rating scale* dan (2) *descriptive graphic rating scale*

Numerical Rating Scale

Komponen pada *numerical rating scale* adalah pernyataan tentang karakteristik atau kualitas tertentu dari sesuatu yang diukur keberadaannya, yang diikuti oleh bilangan yang menunjukkan kualitas keberadaan tersebut.

Contoh 6.1

Berikut ini adalah contoh ***numerical rating scale*** yang mengukur tingkat partisipasi siswa dalam diskusi kelompok yang dapat diisi oleh guru atau pengamat.

Petunjuk:
Nyatakan tingkatan dari setiap pernyataan atau jawaban dari pertanyaan berikut ini dengan cara melingkari salah satu bilangan yang ada di depan pernyataan atau pertanyaan tersebut. Bilangan-bilangan itu mengandung makna:

1 = tidak memuaskan
2 = di bawah rata-rata
3 = rata-rata
4 = di atas rata-rata
5 = sempurna

Nama Siswa yang Diamati: ______________

1.	Seberapa aktifkah siswa berpartisipasi dalam kegiatan diskusi?	1 2 3 4 5
2.	Seberapa baikkah jalinan hubungan baik antara siswa tersebut dengan kelompoknya?	1 2 3 4 5
3.	Seberapa besar kontribusi siswa tersebut dalam pemecahan persoalan yang muncul dalam diskusi?	1 2 3 4 5
4.	dst	1 2 3 4 5

Perhatikanlah bahwa rating dari 1 sampai dengan 5 tersebut dapat dimodifikasi menjadi 4 skala, misalnya 1 = kurang, 2 = cukup, 3 = bagus, dan 4 = bagus sekali, atau menjadi 3 skala, misalnya 1 = kurang, 2 = cukup, dan 3 = bagus.

Numerical rating scale dapat saja dipakai untuk mengukur kemampuan seseorang dalam kegiatan tertentu yang terkait dengan aspek psikomotor.

Contoh 6.2

Berikut ini adalah contoh *numerical rating scale* untuk penggunaan termometer air raksa yang dapat diisi oleh guru atau pengamat.

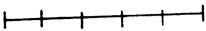
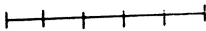
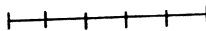
Petunjuk:
Nyatakan tingkatan dari setiap pernyataan atau jawaban dari pertanyaan berikut ini dengan memberi tanda centang (√) pada kolom yang tepat yang ada di depan pernyataan atau pertanyaan tersebut. Angka-angka itu mengandung makna:

5 = sangat tepat
4 = tepat
3 = agak tepat
2 = tidak tepat
1 = sangat tidak tepat

Nama Siswa yang Diamati: ____________

No	Indikator	Jawaban				
		1	2	3	4	5
1	Cara mengeluarkan termometer dari tempatnya.					
2	Cara menurunkan air raksa					
3	Cara memasang termometer pada orang yang diukur suhunya.					
4	Cara mengambil termometer dari tubuh orang yang diukur suhunya					
5	Cara membaca tinggi air raksa dalam pipa kapiler termometer					

Descriptive Graphic Rating Scale

Tipe *rating scale* ini hampir sama dengan *numerical rating scale*. Bedanya adalah bahwa kualitas sesuatu yang dikerjakan digambarkan dalam suatu kontinum pada suatu garis.

Contoh 6.3

Berikut ini adalah contoh *graphic numercal rating scale* yang terkait dengan keaktifan siswa dalam diskusi kelompok yang dapat diisi oleh guru atau pengamat.

Petunjuk:
Nyatakan tingkatan dari setiap pernyataan atau jawaban dari pertanyaan berikut ini dengan cara memberi tanda centang (√) pada tempat yang sesuai.

Nama Siswa yang Diamati: ____________

1. Seberapa aktifkah siswa berpartisipasi dalam kegiatan diskusi?	Sangat Aktif	├──┼──┼──┼──┼──┤	Sangat Tidak Aktif
2. Seberapa baikkah jalinan hubungan baik antara siswa tersebut dengan kelompoknya?	Sangat Baik	├──┼──┼──┼──┼──┤	Sangat Tidak Baik
3. Seberapa besar kontribusi siswa tersebut dalam pemecahan persoalan yang muncul dalam diskusi?	Sangat Berarti	├──┼──┼──┼──┼──┤	Sangat Tidak Berarti
4. dst			

SKALA SIKAP

Untuk dapat memahami pengukuran skala sikap, maka harus dimengerti bahwa sikap adalah suatu konsep psikologik yang harus secara jelas dapat dibedakan dengan kontruks psikologik lainnya, seperti kepercayaan diri, minat, dan opini. Konsep mengenai sikap didiskusikan lebih lanjut di Bab VII. Seperti konstruks psikologik lainnya, maka sikap haruslah memenuhi dua kriteria yaitu dapat diamati dan dapat diukur.

Ada beberapa cara untuk mengukur skala sikap, di antaranya adalah Skala Likert, Skala Thurstone, dan skala beda semantik.

Skala Likert

Model pengukuran dengan skala Likert sebenarnya bernama model *summated ratings* (Sumadi Suryabrata, 2000: 183). Namun, karena modelnya pertama kali diusulkan oleh Rensis Likert, maka model *summated ratings* dikenal dengan skala model Likert, disingkat skala Likert. Skala ini pada dasarnya tergolong untuk mengukur sikap, yang karenanya sering disebut skala sikap.

Prinsip utama skala Likert adalah menentukan lokasi kedudukan seseorang dalam suatu kontinum suatu aspek terhadap suatu objek, mulai dari sangat negatif sampai dengan sangat positif. Penentuan lokasi itu dilakukan dengan mengkuantifikasi pendapat seseorang terhadap pertanyaan atau pernyataan yang disediakan. Pengkuantifikasian itu untuk menunjukkan intensitas sikap yang diukur.

Untuk skala Likert digunakan skala lima dengan 1 (satu) berarti sangat negatif dan skala 5 (lima) berarti sangat positif. Bentuk pernyataan sangat negatif dapat diganti sangat tidak setuju, sangat tidak baik, sangat tidak menarik, dan semacamnya tergantung aspek apa yang dipersoalkan. Kadang-kadang, skala yang di tengah yaitu 3 (netral) dihilangkan, sehingga hanya terdapat, misalnya 1 (sangat tidak setuju), 2 (tidak setuju), 4 (setuju), dan 5 (sangat setuju). Namun sebenarnya tindakan menghilangkan *rating* yang di tengah dapat dianggap sebagai tindakan yang mengingkari kenyataan, sebab pada hakikatnya dalam kehidupan sehari-hari keadaan yang di tengah itu ada, misalnya dalam Pemilu Presiden, seseorang tidak menyatakan pendapatnya mengenai siapa yang pantas menjabat presiden.

Terkait dengan ini, Permendikbud Nomor 104 Tahun 2014, menyederhanakan urutannya menjadi 4 kelompok, yaitu: 4 – selalu, 3 – sering, 2 – kadang-kadang, dan 1 – tidak pernah. Kadang-kadang membedakannya menjadi 4 – sangat baik (SB), 3 – baik (B), 2 – sukup (C), dan 1 – kurang (K). Tentu saja respons untuk skala Likert bisa bermacam-macam bentuknya, tergantung pendapat penilai itu sendiri. Masing-masing mempunyai keunggulan dan kelemahan sendiri-sendiri.

Secara umum, Gable (1986: 42) membedakan menjadi 5 macam tingkatan (*rating*), yaitu: (1) *rating agreement*, (2) *rating frequency*, (3) *rating importance*, (4) *rating quality*, dan (5) *rating likelihood*.

Contoh *rating agreement* adalah: *strongly aggree*, *agree*, *undecided*, *disagree*, dan *strongly disagree*. Contoh *rating frequency* adalah: *always*, *usually*, *about half the time*, *seldom*, dan *never*. Contoh *rating importance* adalah *very important*, *important*, *moderately important*, *of little important*, dan *unimportat*. Contoh *rating quality* adalah: *excellent*, *above average*, *average*, *below average*, dan *extremely poor*. Contoh rating *likelihood* adalah: *always true*, *often true*, *occasionally true*, *usually not true*, *dan almost never true*.

Skala Likert biasanya diisi oleh responden (seseorang yang dikenai angket) berdasarkan pendapatnya sendiri yang harapannya tanpa dipengaruhi oleh orang lain.

Contoh 6.4

Berikut ini adalah contoh skala Likert dengan menggunakan pilihan sangat setuju, setuju, tidak mempunyai pendapat, tidak setuju, dan sangat tidak setuju.

Petunjuk:
Jawablah semua butir di bawah ini dengan memberi tanda cek (√) pada tempat yang tersedia sesuai dengan keyakinan Anda!

1. Matematika sangat berguna dalam kehidupan sehari-hari.

sangat setuju	setuju	tidak mempunyai pendapat	tidak setuju	sangat tidak setuju

2. Untuk mendapat nilai yang tinggi pada mata pelajaran matematika, saya harus bekerja keras.

sangat setuju	setuju	tidak mempunyai pendapat	tidak setuju	sangat tidak setuju

3. Saya harus memperhatikan dengan serius saat guru berbicara di depan kelas.

sangat setuju	setuju	tidak mempunyai pendapat	tidak setuju	sangat tidak setuju

4. Saya tidak perlu belajar keras, karena guru akan memberi nilai baik kepada saya.

sangat setuju	setuju	tidak mempunyai pendapat	tidak setuju	sangat tidak setuju

5. Saya belajar matematika karena terpaksa.

sangat setuju	setuju	tidak mempunyai pendapat	tidak setuju	sangat tidak setuju

Contoh 6.5

Kadang-kadang skala sikap pada Contoh 6.4 dinyatakan dalam bentuk yang lebih kompak seperti di bawah ini. Hal itu dilakukan untuk menghemat kertas.

Jawablah semua butir di bawah ini dengan memberi tanda cek (√) pada tempat yang tersedia sesuai dengan keyakinan Anda!

Keterangan: SS = sangat setuju
S = setuju

TMP = tidak mempunyai pendapat
TS = tidak setuju
STS = sangat tidan setuju

No	Pernyataan	SS	S	TMP	TS	STS
1	Matematika sangat berguna dalam kehidupan sehari-hari.					
2	Untuk mendapat nilai yang tinggi pada mata pelajaran matematika, saya harus bekerja keras.					
3	Saya harus memperhatikan dengan serius saat guru berbicara di depan kelas.					
4	Saya tidak perlu belajar keras, karena guru akan memberi nilai baik kepada saya.					

Kadang-kadang skala Likert dipakai untuk mengukur pernyataan kognitif, yaitu pernyataan tingkah laku yang berkenaan dengan suatu objek sikap tertentu. Ada dua macam pernyataan kognitif. Pertama, pernyataan yang menyatakan apa yang akan dilakukan terhadap suatu objek sikap tertentu. Misalnya: Bila saya boleh memilih maka saya akan membeli kendaraan bermesin diesel. Kedua, pernyataan yang menyatakan kecenderungan tindakan sosial. Misalnya: Pemerintah seharusnya meringankan pajak bagi kendaraan bermesin diesel.

Skala Thurstone

Model pengukuran skala Thorstone dikembangkan pertama kali oleh Louis Thurstone (Sumadi Suryabrata, 2000: 200). Thurstone oleh para ahli ilmu-ilmu sosial dianggap "bapak" penyusunan skala untuk mengukur sikap.

Skala Thurstone mirip dengan skala Likert, namun biasanya rentangan skala pada skala Thurtone lebih lebar, berkisar antara 7 sampai dengan 11 skala. Pada skala Thurstone, responden juga hanya membubuhkan tanda cek (√) pada tempat yang disediakan. Berikut ini adalah contoh skala Thurstone.

Contoh 6.7

Berikut ini adalah contoh skala Thurstone untuk mengukur sikap siswa terhadap matematika dan pembelajarannya.

Petunjuk:
Berilah tanda cek (√) pada tempat yang disediakan. Skala 7 menunjukkan sangat setuju, sedangkan skala 1 menunjukkan sangat tidak setuju.

No	Pernyataan	1	2	3	4	5	6	7
1	Matematika sangat berguna dalam kehidupan sehari-hari.							
2	Untuk mendapat nilai yang tinggi pada mata pelajaran matematika, saya harus bekerja keras.							
3	Saya harus memperhatikan dengan serius saat guru berbicara di depan kelas.							
4	Saya tidak perlu belajar keras, karena guru akan memberi nilai baik kepada saya.							
5	Saya belajar matematika karena terpaksa.							

Skala Beda Semantik

Skala beda semantik mirip dengan skala Thurstone, namun pada skala beda semantik, seseorang diminta pendapatnya untuk suatu hal dari berbagai sudut pandang yang berbeda. Berikut ini contoh skala beda semantik.

Contoh 6.8

Berikut ini adalah skala beda semantik untuk mengukur sikap siswa terhadap mata pelajaran yang diikutinya.

Petunjuk: Berilah tanda cek (√) pada tempat yang disediakan sesuai dengan perasaan dan atau pendapat Anda!

1. Mata pelajaran Matematika:

Menyenangkan								Membosankan
Sulit								Mudah
Bermanfaat								Sia-sia
Menantang								Menjemukan
Hafalan								Penalaran

2. Mata pelajaran Sejarah:

Menyenangkan								Membosankan
Sulit								Mudah
Bermanfaat								Sia-sia
Menantang								Menjemukan
Hafalan								Penalaran

Perlu diketahui bahwa harus digunakan kata sifat pada skala beda semantik, misalnya menyenangkan-membosankan dan sulit-mudah. Tidak boleh digunakan kata tidak, misalnya menyenangkan-tidak menyenangkan. sulit-tidak sulit.

Angket

Kadang-kadang instrumen non tes dibuat dalam bentuk yang menyerupai angket, walaupun pada dasarnya angket tersebut kalau ditelusur lebih dalam sebenarnya adalah skala Likert atau skala Thurstone. Model angket ini dipilih karena bisa lebih luwes karena *option* (pilihan jawaban) bisa beraneka ragam.

Contoh 6.9

Berikut ini adalah angket untuk mengukur motivasi siswa dalam pembelajaran matematika.

Petunjuk:
Jawablah semua butir soal di bawah ini dengan melingkari jawaban yang paing tepat sesuai dengan kondisi Anda!

1. Matematika sangat berguna dalam kehidupan sehari-hari.
   a. sangat setuju
   b. setuju
   c. tidak mempunyai pendapat
   d. tidak setuju
   e. sangat tidak setuju
2. Untuk mendapat nilai yang tinggi pada mata pelajaran matematika, saya harus bekerja keras.
   a. sangat setuju
   b. setuju
   c. tidak mempunyai pendapat
   d. tidak setuju
   e. sangat tidak setuju
3. Saya harus memperhatikan dengan serius saat guru berbicara di depan kelas.
   a. sangat setuju
   b. setuju
   c. tidak mempunyai pendapat
   d. tidak setuju
   e. sangat tidak setuju
4. Saya tidak perlu belajar keras, karena guru akan memberi nilai baik kepada saya.
   a. sangat setuju
   b. setuju
   c. tidak mempunyai pendapat
   d. tidak setuju
   e. sangat tidak setuju

5. Saya belajar matematika karena terpaksa.
   a. sangat setuju
   b. setuju
   c. tidak mempunyai pendapat
   d. tidak setuju
   e. sangat tidak setuju

BAHAN DISKUSI

1. Dikaitkan dengan taksonomi Bloom, non-tes lebih tepat dipakai untuk mengukur hasil pembelajaran ranah kognitif, ranah afektif, atau ranah psikomotor? Mengapa?

2. Semula, ada lima tingkatan pada skala Likert, yaitu SS = sangat setuju, S = setuju, TMP = tidak mempunyai pendapat. TS = tidak setuju, dan STS = sangat tidak setuju. Setujukah Anda kalau tingkatan TMP dihilangkan? Mengapa?

3. a. Kajilah teori mengenai motivasi berprestasi.
   b. Berdasarkan teori tersebut, tulislah indikator pengukurannya.
   c. Berdasarkan indikator tersebut, buatlah skala Likert untuk mengukur motivasi seseorang.

4. a Kajilah teori mengenai berbagai jenis pembelajaran kooperatif.
   b. Berdasarkan hal tersebut, buatlah skala beda semantik yang menanyakan kepada guru mengenai pendapatnya mengenai berbagai jenis pembelajaran kooperatif tersebut.

5. a Kajilah teori mengenai cara *start* pada lari cepat 100 meter.
   b. Berdasatkan itu buatlah *numerical rating* untuk melihat cara pelari melakukan start.

****

BAB VII
PENILAIAN RANAH AFEKTIF

PENDAHULUAN

Kebanyakan pendidik menganggap bahwa hasil belajar ranah afektif tidaklah penting. Mereka mengganggap bahwa yang terpenting adalah hasil belajar yang lebih bersifat kognitif (seperti matematika) atau psikomotorik (seperti olah raga dan menari). Para pendidik lupa bahwa kadar afektif seseorang akan menentukan kehidupan seseorang di masa mendatang. Sebagai contoh, jika seseorang percaya bahwa kesehatan adalah penting, maka mereka akan berusaha untuk memelihara kesehatannya sepanjang masa. Jika seseorang percaya matematika berguna di masa depan dan dia percaya bahwa dia dapat mempelajari matematika dengan baik, maka seseorang akan terus berusaha untuk belajar matematika. Sebaliknya, jika seorang siswa percaya bahwa matematika tidak berguna, maka dia akan tidak dengan sungguh-sungguh mengikuti pembelajaran matematika di kelas. Dengan demikian, para pendidik bertugas untuk selalu meningkatkan kadar afektif para peserta didiknya terkait dengan mata kuliah atau mata pelajaran yang diampunya, atau paling tidak menjaga agar kadar afektif peserta didiknya tidak menurun. Menurunnya kadar afektif peserta didik, menandakan bahwa peserta didik tidak tertarik terhadap mata kuliah atau mata pelajaran tersebut.

Dengan dapat diukurnya kadar afektif peserta didik secara kontinu, pendidik dapat pula melakukan refleksi atas proses pembelajarannya. Jika kadar afektif peserta didiknya cenderung menurun, maka terdapat indikasi bahwa proses pembelajaran yang telah berlangsung kurang menarik, sehingga pendidik dapat melakukan perbaikan proses pembelajaran berikutnya.

Misalnya, setelah dilakukan pengukuran, ternyata tingkat kecemasan sebagian peserta didiknya tinggi. Dalam kasus seperti ini, pendidik dapat mengurangi tingkat kecemasan tersebut. Misalnya dengan menciptakan kelas yang nyaman, memberikan kesempatan kepada peserta didik untuk memperbaiki pekerjaan, dan menyajikan pembelajaran dengan menarik. Jika tingkat motivasi sebagian peserta didiknya rendah, pendidik dapat meningkatkan motivasi dengan berbagai cara, misalnya dengan menggunakan berbagai macam model persentasi.

PENGERTIAN RANAH AFEKTIF

Pengertian afektif menurut Anderson (1981:3) dan Gable (1986:2-4) adalah kualitas yang menunjukkan cara khas seseorang menyatakan perasaan atau mengungkapkan emosinya (*qualities which present people's typical ways of feeling or expressing their emotion*). Ada dua ciri utama ranah afektif. Ciri pertama adalah melibatkan perasaan dan emosi, dan ciri kedua adalah perasaan tersebut memiliki pola ungkapan yang relatif sama dalam berbagai situasi ruang dan waktu. Kedua ciri ranah afektif tersebut memuat tiga komponen afektif, yaitu *intensity* (intensitas), *direction* (arah), dan *target* (sasaran atau objek).

Menurut Anderson (1981:5), ada beberapa ciri ranah afektif, yaitu: (1) ada unsur perasaan, (2) ada pola perasaan, (3) ada tingkatan intensitas perasaan, (4) ada arah perasaan (positif atau negatif), dan (5) ada sasaran, baik sasaran yang diketahui maupun sasaran yang tidak diketahui.

PENGGOLONGAN RANAH AFEKTIF

Anderson (1981: 29) mengatakan bahwa terdapat 7 karakteristik afektif, yaitu: (1) sikap (*attitude*), (2) minat (*interest*), (3) nilai (*value*), (4) pilihan (*preference*), (5) kepercayaan diri akademik (*academic self esteem*), (6) lokus kendali (*locus of control*), dan (7) kecemasan (*anxiety*).

Seperti telah disebutkan di muka, setiap karakteristik mempunyai intensitas, arah, dan sasaran. Intensitas adalah ukuran derajat atau kekuatan perasaan, arah adalah sifat yang menyatakan apakah perasaan itu positif, netral, atau negatif, sedangkan sasaran adalah objek, perilaku, atau gagasan yang dituju oleh arah perasaan itu. Kecuali karakteristik tersebut, beberapa pakar juga memasukkan motivasi ke dalam ranah afektif (Djemari Mardapi, dkk, 2002: 33; Suryanto, 2001: 49).

Sikap (*attitude*) diartikan sebagai kecenderungan untuk merespon secara positif (*favorable*) atau secara negatif (*unfavorable*) terhadap suatu objek (Anderson, 1981: 29). Ini berarti sikap adalah kecenderungan seseorang untuk menanggapi suatu objek dalam tanggapan suka (sikap positif) atau tidak suka (sikap negatif). Adanya sikap positif seseorang terhadap

suatu objek menunjukkan bahwa seseorang tersebut menyenangi dan atau menghargai objek tersebut, sedangkan adanya sikap negatif seseorang terhadap suatu objek menunjukkan bahwa seseorang tersebut tidak menyenangi atau tidak menghargai objek tersebut. Kata-kata yang dapat digunakan untuk mengukur sikap, antara lain, menyenangi – tidak menyenangi, diingini – dibenci, menerima – menolak, dan tertarik – tidak tertarik. Dalam pembelajaran matematika, misalnya, dapat diukur sikap siswa terhadap buku matematika, belajar matematika, pengerjaan soal matematika, dan guru matematika.

Minat (*interest*) diartikan sebagai watak yang terorganisir melalui pengalaman yang mendorong seseorang untuk mendalami suatu objek, pengertian, keterampilan, atau tujuan untuk mendapatkan suatu kemahiran atau penguasaan tertentu (Anderson, 1981: 30). Dalam pembelajaran matematika, misalnya, dapat diukur minat siswa untuk mengikuti pelajaran matematika, mempelajari tokoh-tokoh matematika, dan menggunakan matematika di luar kelas.

Nilai (*value*) diartikan sebagai objek, aktivitas, atau pandangan yang diapresiasi oleh seseorang dalam mengarahkan minat, sikap, atau kepuasannya (Anderson, 1981: 31). Dalam pembelajaran matematika, misalnya, dapat diukur pandangan siswa terhadap guru matematika dan penggunaan matematika. Misalnya siswa memandang penting belajar matematika, maka nilai mereka terhadap matematika tinggi.

Pilihan (*preference*) adalah kecenderungan untuk memilih suatu objek, aktivitas, atau gagasan dibandingkan dengan objek, aktivitas, atau gagasan lain (Anderson, 1981: 32). Pilihan melibatkan pemilihan di antara dua objek, dua aktivitas, atau dua gagasan atau lebih. Oleh karena itu, biasanya pilihan bersifat relatif, misalnya lebih menyenangi ini daripada itu, lebih suka menjadi itu daripada ini. Dalam pembelajaran matematika, misalnya, dapat diukur pilihan siswa terhadap berbagai hal, misalnya antara mempelajari matematika dibandingkan dengan mata pelajaran lain dan antara menjadi matematikawan atau menjadi dokter.

Konsep diri (*self-esteem*) diartikan sebagai persepsi seseorang terhadap dirinya sendiri (Anderson, 1981: 32). Menurut Smith (Tim Pascasarjana UNY, 2003b: 10), konsep diri adalah evaluasi yang dilakukan seseorang terhadap kelemahan yang dimilikinya. Dalam pembelajaran matematika, misalnya, konsep diri siswa dapat diukur melalui kepercayaannya dalam mempelajari matematika atau bagian-bagiannya, kepercayaannya dalam mengharapkan pekerjaan kelak jika menguasai matematika, dan kepercayaannya dalam menyelesaikan soal-soal matematika.

Lokus kendali (*locus of control*) adalah seberapa jauh seseorang dapat menerima sesuatu karena tindakannya atau konsekuensi dari tindakannya (Anderson, 1981: 33). Seseorang dengan lokus kendali internal adalah orang

yang percaya bahwa berhasil atau gagal adalah karena usahanya sendiri. Seseorang dengan lokus kendali eksternal cenderung lebih yakin bahwa faktor lain, seperti kemujuran atau tindakan orang lain, yang menyebabkan berhasil atau gagal. Dalam konteks ini, seseorang yang yakin bahwa keberhasilan di sekolah karena kemujuran atau faktor lainnya cenderung untuk tidak mau bekerja keras. Di sisi lain, siswa yakin bahwa keberhasilan atau kegagalan terutama dikarenakan usahanya sendiri dapat diharapkan untuk mau bekerja keras. Dalam konteks pembelajaran, lokus kendali dapat diukur dari seberapa jauh seorang siswa percaya bahwa apa yang diperolehnya (misalnya nilai untuk mata pelajaran tertentu) adalah karena usahanya sendiri atau karena faktor-faktor lain di luar dirinya.

Kecemasan (*anxiety*) diartikan sebagai pengalaman mendapatkan tekanan yang menghasilkan ancaman kepada seseorang, baik secara riil maupun secara imajiner (Anderson, 1981:34). Senada dengan itu, Hall, Lindsay, dan Campbell (1970) mengatakan bahwa kecemasan adalah pengalaman menegangkan sebagai akibat dari ketakutan, baik ketakutan karena sesuatu yang bersifat nyata atau bersifat imajinatif. Dalam pembelajaran matematika, misalnya, dapat diukur kecemasan seseorang menempuh tes matematika, kecemasan mengerjakan tugas matematika, dan kecemasan seseorang menghadapi guru matematika.

Beberapa pakar memasukkan motivasi ke dalam ranah afektif. Motivasi adalah proses internal yang mengaktifkan, membimbing, dan mempertahankan perilaku dalam suatu rentang waktu tertentu (Muhamad Nur, 1999: 2). Dalam bahasa sederhana, motivasi adalah apa yang membuat seseorang berbuat, membuat seseorang untuk tetap berbuat, dan menentukan ke arah mana seseorang akan perbuat. Motivasi dapat bervariasi dalam intensitas dan arah. Motivasi tidak hanya penting untuk menjadikan siswa terlibat dalam kegiatan akademik, tetapi juga penting dalam menentukan seberapa jauh siswa akan belajar dari suatu kegiatan pembelajaran atau seberapa jauh menyerap informasi yang disajikan kepada mereka.

Menurut Krathwol (dalam Reynolds, Livingstone, dan Willson, 2010: 175), ada lima tingkatan ranah afektif, yaitu: (1) menerima (*receiving* atau *attending)*, (2) merespons (*responding*), (3) menilai (*valuing*), (4) mengelola (*organization*), dan (5) memjadi karakter (*characterization*).

Receiving merupakan keinginan siswa untuk memperhatikan fenomena atau stimuli tertentu, misalnya kegiatan di kelas, buku-buku, dan musik. Dari sudut pandang pendidik, *receiving* berkenaan dengan upaya untuk mendapatkan dan mengarahkan perhatian peserta didik agar dapat mengikuti pembelajaran dengan baik. Tugas pendidik adalah mengarahkan perhatian peserta didik pada fenomena yang menjadi objek pembelajaran.

Responding merupakan partisipasi aktif peserta didik. Pada tingkatan ini, peserta didik tidak saja memperhatikan fenomena tertentu yang muncul, tetapi juga sudah memberikan respons dalam berbagai cara. Hasil pembelajaran pada tingkatan ini adalah pemerolehan respons, keinginan untuk memberikan respons, dan kepuasan dalam memberikan respons. Pada tingkatan ini muncul minat, yaitu hal-hal yang menekankan kepada pencarian hasil dan kepuasan pada aktivitas tertentu.

Valuing berkenaan dengan penentuan nilai atau *worth* yang dilekatkan oleh siswa kepada objek, fenomena, atau *behavior* tertentu. Rentangan dari *valuing* ini mulai dari penerimaan suatu nilai (yang dimaksudkan untuk meningkatkan keterampilan) sampai dengan komitmen yang tinggi terhadap sesuatu. Hasil belajar pada tingkatan ini berkenaan dengan perilaku yang konsisten dan stabil untuk membuat nilai. Pada tingkatan ini, muncul *attitudes* (sikap) dan *appreciation* (apresiasi).

Organization berkaitan dengan pengumpulan nilai-nilai yang berbeda dalam satu kaitan, menyelesaikan konflik yang ada, dan mulai membangun sistem nilai internal yang konsisten. Hasil pembelajaran pada tingkatan ini adalah konseptualisasi nilai atau organisasi sistem nilai.

Pada tingkatan *characterization*, seseorang telah memiliki sistem nilai yang mengendalikan perilakunya sehingga terbentuk gaya hidup (*life style*). Gaya hidup ini akan bertahan lama dan sulit untuk diubah.

INSTRUMEN PENILAIAN RANAH AFEKTIF

Paling tidak ada 3 model instrumen untuk mengukur ranah afektif, yaitu: skala Likert, skala Thurstone, dan skala beda semantik (*semantic differential scale*). Demi kemudahan, kadang-kadang skala-skala tersebut dibuat dalam bentuk yang menyerupai kuesioner atau angket.

Skala Likert, skala Thurstone, skala beda semantik, dan skala Likert dalam bentuk angket telah dibicarakan pada Bab VI.

Contoh 7.1

Berikut diberikan lagi contoh skala Likert mengenai sikap siswa terhadap matematika yang ada di Bab VI.

Petunjuk:

Perhatikan pernyataan-pernyataan di bawah ini. Untuk masing-masing pernyataan, berikan pendapat Anda dengan memberi centang pada jawaban SS, S, TMP, TS dan STS dengan penjelasan sebagai berikut

SS = sangat setuju
S = setuju
TMP = tidak mempunyai pendapat
TS = tidak setuju
STS = sangat tidak setuju

No	Pernyataan	Jawaban				
		SS	S	TMP	TS	STS
1	Matematika sangat berguna dalam kehidupan sehari-hari.					
2	Untuk mendapat nilai yang tinggi pada mata pelajaran matematika, saya harus belajar dengan sungguh-sungguh.					
3	Saya harus memperhatikan dengan serius saat guru matematika berbicara di depan kelas.					
4	Saya tidak perlu belajar keras, karena guru akan memberi nilai baik kepada saya.					
5	Saya belajar matematika karena terpaksa.					

Perhatikan tiga butir pertama dari skala Likert pada Contoh 7.1. Pada tiga butir pertama tersebut, jika responden memilih SS, maka dia mendapat skor 5; jika responden memilih S, maka dia mendapat skor 4; jika responden memilih TMP, maka dia mendapat skor 3; jika responden memilih TS, maka dia mendapat skor 2; dan jika responden memilih STS, maka dia mendapat skor 1.

Di sisi lain, untuk butir nomor empat dan lima, jika responden memilih SS, maka dia mendapat skor 1; jika responden memilih S, maka dia mendapat skor 2; jika responden memilih TMP, maka dia mendapat skor 3; jika responden memilih TS, maka dia mendapat skor 4; dan jika responden memilih STS, maka dia mendapat skor 5.

Ini berarti, dua kelompok butir instrumen tersebut mempunyai arah yang berlawanan. Tiga butir yang pertama dikatakan butir instrumen yang mempunyai arah positif, sedangkan dua butir yang terakhir dikatakan mempunyai arah negatif.

Pada Kurikulum 2013, seperti yang tertuang pada Permendikbud 104 Tahun 2014 tentang Penilaian Hasil Belajar, penskoran sikap menggunakan skala empat, yaitu: 4 = sangat baik, 3 = baik, 2 = cukup, dan 1 = kurang; atau 4 = selalu, 3 = sering, 2 = jarang, dan 1 = sangat jarang.

PENGEMBANGAN INSTRUMEN RANAH AFEKTIF

Pada dasarnya pengembangan alat ukur ranah afektif. mengikuti langkah-langkah yang telah dikemukakan pada Bab IV.

Secara ringkas, seperti halnya penyusunan instrumen pada ranah kognitif, langkah-langkah penyusunan instrumen (termasuk untuk ranah afektif) adalah: (Djemari Mardapi, dkk, 2002: 20) (1) menyusun spesifikasi instrumen, (2) menulis butir-butir instrumen, (3) menelaah butir-butir instrumen, (4) melakukan uji coba, (5) menganalisis butir instrumen berdasar uji coba, (6) melakukan revisi terhadap butir-butir instrumen yang kurang baik, jika memungkinkan, (7) merakit instrumen dengan menetapkan butir-butir yang dipakai, (8) melaksanakan pengukuran (pengujian) pada subjek yang dikehendaki, (9) menafsirkan hasil yang diperoleh.

Pada bagian ini dicontohkan pengembangan spesifikasi dan butir-butir instrumen untuk sikap yang dinyatakan dalam bentuk angket.

Contoh 7.2

Berikut ini adalah contoh spesifikasi instrumen untuk sikap.

a. Tujuan: untuk mengukur sikap siswa terhadap matematika dan pembelajarannya di kelas.

b. Kisi-kisi:
   1) Definisi konseptual: Sikap terhadap matematika dan pembelajarannya di kelas adalah kecenderungan untuk merespon secara positif (*favorable*) atau secara negatif (*unfavorable*) terhadap matematika dan pembelajarannya di kelas.
   2) Definisi operasional: Sikap terhadap matematika dan pembelajarannya di kelas adalah kecenderungan untuk memberikan pendapat mengenai kegunaan matematika, cara guru membuka pembelajaran, media pembelajaran yang digunakan guru, interaksi guru dan siswa, dan cara pemberian umpan balik kepada siswa.
   3) Indikator/Deskriptor:
      (1) sikap siswa terhadap kegunaan matematika
      (2) sikap siswa terhadap cara guru membuka pembelajaran
      (3) sikap siswa terhadap media yang digunakan guru
      (4) sikap siswa terhadap interaksi guru dan siswa di kelas
      (5) sikap siswa terhadap cara pemberian umpan balik kepada siswa
   4) Jenis instrumen: skala Likert (dalam bentuk angket)
   5) Banyaknya butir dan nomor butir:

No	Indikator/Deskriptor	Nomor Butir (arah positif)	Nomor Butir (arah negatif)	Banyaknya Butir
1	Sikap siswa terhadap kegunaan matematika	1	8	2
2	Sikap siswa terhadap cara guru membuka pembelajaran	2	9	2
3	Sikap siswa terhadap media yang digunakan guru	3, 4	10	3
4	sikap siswa terhadap interaksi guru dan siswa di kelas	5	11, 12	3
5	sikap siswa terhadap cara pemberian umpan balik kepada siswa	6, 7	13, 14, 15	5

Perhatikan bahwa banyaknya butir dengan arah positif dan dengan arah negatif hampir seimbang. Hal ini diperlukan agar para responden yang dikenai instrumen membaca dengan sungguh-sungguh pernyataannya. Perhatikan juga bahwa ada minimal dua butir instrumen pada setiap indikator. Hal ini diperlukan, karena pada uji coba bisa saja butir instrumen gugur karena tidak memenuhi persyaratan.

Setelah dibuat spesifikasi, ditulis butir-butir untuk spesifikasi instrumen pada Contoh 7,3.

Contoh 7.3

Berikut ini adalah contoh butir instrumen untuk mengukur skala sikap berdasarkan spesifikasi pada Contoh 7.2.

1. Menurut Anda, apakah seseorang perlu menguasai matematika untuk bekal kehidupan di masa depan?
   a. sangat perlu
   b. perlu
   c. netral
   d. tidak perlu
   d. sangat tidak perlu

2. Menurut pendapat Anda, bagaimana cara guru dalam memulai pembelajaran?
   a. sangat menarik
   b. cukup menarik
   c. tidak mempunyai pendapat
   d. tidak menarik
   e. sangat tidak menarik

3. Menurut pendapat Anda, bagaimana kualitas media yang digunakan guru?
   a. sangat bagus
   b. cukup bagus
   c. tidak mempunyai pendapat
   c. tidak bagus
   d sama sekali tidak bagus

PENSKORAN INSTRUMEN PENILAIAN RANAH AFEKTIF

Sistem penskoran instrumen afektif tergantung kepada jenis instrumen yang dipakai. Jika instrumen yang dipakai adalah skala Likert dengan 5 skala (SS, S, N, TS, STS), maka skor tertinggi setiap butir adalah 5 dan skor terrendah adalah 1. Jika digunakan skala Thurstone dengan 7 skala, maka skor tertinggi setiap butir adalah 7 dan skor terrendah adalah 1. Jika digunakan angket dengan 4 pilihan, maka skor tertinggi untuk setiap butir adalah 4 dan skor terrendah adalah 1. Kurikulum 2013 menggunakan 4 skala, yaitu 4 = sangat baik, 3 = baik, 2 = cukup, dam 1 = kurang.

Dengan menjumlahkan skor untuk seluruh butir, maka diperoleh skor total yang merupakan skor untuk ranah afektif. Jika banyaknya butir cukup banyak, skor total ini dapat dianggap skor suatu variabel yang berskala interval, walaupun skor untuk masing-masing butir merupakan skor dengan skala ordinal.

PENAFSIRAN HASIL PENGUKURAN RANAH AFEKTIF

Biasanya nilai untuk ranah afektif tidak dalam bentuk kuantitatif, tetapi dalam bentuk kualitatif, misalnya sangat positif (sangat tinggi), positif (tinggi), cukup, negatif (rendah), dan sangat negatif (sangat rendah). Oleh karena itu, diperlukan aturan untuk mengkonversi dari skor mentah ke nilai hasil pembelajaran. Aturan itu dapat mengacu kepada patokan tertentu atau dengan menggunakan acuan norma.

Misalnya skor terrendah yang dicapai siswa adalah 5 dan skor tertinggi adalah 100, maka dengan acuan patokan, dapat dilakukan konversi dengan aturan berikut:

Tabel 7.1. Aturan Konversi Skor dengan Penilaian Acuan Patokan (PAP)

Skor	Nilai
$5 \leq$ skor < 20	Sangat rendah
$21 \leq$ skor < 40	Rendah
$41 \leq$ skor < 60	Cukup
$61 \leq$ skor < 80	Tinggi
$81 \leq$ skor ≤ 100	Sangat tinggi

Jika digunakan acuan norma, maka diperlukan penghitungan rataan kelas dan simpangan baku kelas. Misalnya rataan kelasnya adalah $\overline{X}$ dengan simpangan baku s, maka kita dapat dilakukan konversi dengan aturan berikut.

Tabel 7.2. Aturan Konversi Skor dengan Penilaian Norma (PAN)

Skor	Nilai
skor < $\overline{X} - 1{,}5s$	Sangat rendah
$\overline{X} - 1.5s \le$ skor < $\overline{X} - 0{,}5s$	Rendah
$\overline{X} - 0.5s \le$ skor < $\overline{X} + 0{,}5s$	Cukup
$\overline{X} + 0.5s \le$ skor < $\overline{X} + 1{,}5s$	Tinggi
skor > $\overline{X} + 1.5s$	Sangat tinggi

Kadang-kadang hanya dibedakan menjadi tiga tingkatan, yaitu kurang, sedang, dan bagus atau empat tingkatan yaitu sangat baik, baik, cukup dan kurang. Jika dinyatakan dalam tiga atau empat skala, maka perlu dilakukan modifikasi terhadap Tabel 7.1 dan Tabel 7.2 di atas.

VALIDITAS INSTRUMEN RANAH AFEKTIF

Seperti dibicarakan di Bab II, instrumen penilaian ranah afektif dapat divalidasi melalui tiga cara, yaitu dengan validasi isi (ahli), validasi berdasar kriteria, dan validasi konstruks.

ANALISIS BUTIR PADA INSTRUMEN RANAH AFEKTIF

Setelah proses validasi (isi) dilakukan, maka untuk memilih butir-butir yang baik, maka instrumen tersebut harus dicobakan kepada sekelompok subjek yang kira-kira mempunyai ciri yang sama dengan kelompok subjek yang akan dikenai instrumen. Banyaknya subjek yang dikenai uji coba sekitar 6 – 10 kali banyaknya butir instrumen (Gable, 1986:39).

Sebuah instrumen tentu terdiri dari sejumlah butir-butir instrumen. Skor suatu butir tersebut seharusnya menunjukkan kecenderungan yang sama dengan skor totalnya, yang dalam hal ini diasumsikan bahwa skor total mewakili skor konstruks yang diukur. Dalam konteks ini, skor total instrumen dianggap mewakili karakteristik afektif yang dimaksudkan untuk diukur. Ini berarti harus ada korelasi positif antara skor suatu butir tersebut dengan skor totalnya.

Koefisien korelasi antara skor suatu butir dengan skor total disebut sebagai indeks konsistensi internal butir tersebut (oleh Gable, 1986: 39

disebut sebagai *criterion of internal consistency*). Untuk menghitung konsistensi internal untuk butir ke-i, rumus yang digunakan adalah rumus korelasi momen produk dari Karl Pearson berikut.

$$r_{xy} = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}}$$

dengan: r_{xy} = indeks konsistensi internal untuk butir ke-i, n = banyaknya subjek yang dikenai instrumen, X = skor untuk butir ke-i (dari subjek uji coba), dan Y = skor total (dari subjek uji coba).

Tentu saja, jika terdapat n buah butir, maka akan dilakukan penghitungan sebanyak n kali. Jika indeks konsistensi internal untuk butir ke-i kurang dari 0,3 maka butir tersebut harus dibuang.

Jika instrumennya berupa tes hasil belajar, indeks konsistensi internal ini merupakan indeks daya pembeda. Jika untuk angket, indeks konsistensi internal ini disebut pula dengan indeks daya pembeda angket[1].

Beberapa buku memaknai koefisien korelasi antara skor butir dengan skor total sebagai indeks validitas butir. Konsep itu merupakan konsep yang salah, sebab tidak dikenal adanya konsep validitas butir. Yang dikenal adalah konsep mengenai validitas instrumen.

Pada beberapa buku, untuk menentukan apakah butir dibuang atau dipertahankan pada instrumen dilakukan uji signifikansi koefisien korelasi. Cara ini juga tidak tepat, karena pada dasarnya penentuan apakah butir dibuang atau dipertahankan dalam instrumen bukan uji signifikansi.

RELIABILITAS INSTRUMEN RANAH AFEKTIF

Setelah diperoleh butir-butir yang baik, maka butir-butir yang baik tersebut dirakit menjadi sebuah instrumen yang siap untuk digunakan. Namun demikian, sebelum instrumen tersebut digunakan, perlu dilihat reliabilitasnya terlebih dulu.

Reliabilitas menunjuk kepada konsistensi pengukuran jika dilakukan pengukuran berulang-ulang pada individu-individu atau kelompok-kelompok dalam suatu populasi (*AERA, APA, & NCME*, 1999: 25). Ini berarti, keterandalan suatu tes menunjuk kepada besarnya kesalahan pengukuran yang dihasilkan oleh tes tersebut. Semakin besar koefisien keterandalan suatu tes akan semakin kecil kesalahan pengukurannya.

[1] Penulis tidak menggunakan istilah daya pembeda angket, karena daya pembeda didefinisikan sebagai selisih proporsi kelompok pandai dan kelompok tidak pandai dalam menjawab benar butir soal. Dalam angket, misalnya, tidak ada jawaban benar dan jawaban salah, sehingga istilah daya pembeda angket dianggap tidak tepat.

Seperti didiskusikan di Bab II, ada tiga cara mencari koefisien reliabilitas (termasuk untuk instrumen ranah afektif), yaitu: (a) metode satu kali tes, (b) metode tes ulang, dan (c) metode bentuk paralel. Metode mana yang sebaiknya dipakai, tidak ada aturan baku. Namun, biasanya orang akan memilih metode satu kali tes, sebab metode ini mudah dilakukan dan berbiaya murah dibandingkan dengan dua pendekatan yang lainnya.

Untuk metode satu kali tes, biasanya yang digunakan adalah teknik alpha dengan menggunakan rumus Cronbach-Alpha, seperti yang sudah disampaikan pada Bab III.

BAHAN DISKUSI

1. Pada Kurikulum 2013 dinyatakan bahwa ada dua jenis sikap, yaitu sikap spiritual dan sikap sosial. Contoh sikap sosial adalah bekerja-sama, rasa ingin tahu, disiplin, dan peduli lingkungan. Apakah sikap yang dicontohkan oleh Kurikulum 2013 tersebut memenuhi sifat-sifat sikap yang didefinisikan oleh Anderson? Mengapa?
2. Buatlah spesifikasi dan butir-butir instrumen untuk sikap yang dinyatakan dalam skala Likert untuk mengukur:
   a. motivasi berprestasi siswa
   b. kecemasan siswa menghadapi ujian matematika
   c. locus kendali siswa
   d. kepercayaan diri akademik siswa
3. a. Dapatkah butir-butir angket Contoh 7.3 dihitung tingkat kesulitannya? Mengapa?
   b. Dapatkah butir-butir angket Contoh 7.3 dihitung tingkat daya pembedanya ? Mengapa?
4. Dapatkah angket Contoh 7.3 diestimasi koefisien reliabilitasnya dengan rumus KR-20? Mengapa?
5. Manakah yang lebih sulit, mengukur ranah kognitif atau afektif? Mengapa?
6. Misalnya Anda menggunakan angket untuk mengukur tingkat keseriusan siswa dalam mengikuti pelajaran Matematika. Yakinkah Anda kalau siswa menjawab jujur? Mengapa? Apa yang perlu dilakukan untuk meningkatkan tingkat kejujuran siswa dalam mengisi angket? Jelaskan!

*****

BAB VIII
PENILAIAN RANAH PSIKOMOTOR

PENDAHULUAN

Telah disebutkan di depan bahwa aspek psikomotor menitikberatkan kepada hal-hal yang berkaitan dengan cara tindak atau keterampilan gerak otot. Menurut Reynolds, Livingstone, dan Willson (2010: 175) ranah ini berkaitan dengan "*physical activity*" yang biasanya terkait dengan "*physical education, dance, speech, theater, laboratory (e.g. biology and computer science), or carrer-technical classes such as woodworking, electronics, automotive, or metalwork*". Tujuan pembelajaran di ranah psikomotor selalu terkait dengan tujuan di ranah kognitif, sebab "*almost physical activity involves cognitive processes*".

PENGGOLONGAN RANAH PSIKOMOTOR

Menurut Dave (Tim Pascasarjana UNY, 2003a: 2), aspek psikomotor mencakup imitasi, manipulasi, presisi, artikulasi, dan naturalisasi.

Imitasi adalah kemampuan melakukan kegiatan-kegiatan sederhana dan sama persis dengan yang dilihat atau diperhatikan sebelumnya. Manipulasi adalah kemampuan melakukan kegiatan sederhana berdasarkan pedoman yang disediakan dan belum pernah dilihatnya. Presisi adalah kemampuan melakukan kegiatan secara akurat sehingga mampu menghasilkan produk yang mempunyai tingkat presisi tinggi. Artikulasi adalah kemampuan melakukan kegiatan yang kompleks dengan presisi tinggi, sehingga menghasilkan produk kerja yang utuh. Naturalisasi adalah kemampuan melakukan kegiatan secara refleks.

Di sisi lain, Simpson (Permendikbud Nomor 104 Tahun 2014) mengatakan bahwa ada 7 tingkatan ranah psikomotor, yaitu: (1) persepsi

(*perception*), (2) kesiapan (*set*), (3) meniru (*guided response*), (4) pembiasaan gerakan (mechanism), (5) mahir (*complex or overt response*), (6) adaptasi (*adaption*), dan (7) menjadi tindakan orisinal (origination).

Pada tingkatan persepsi, seseorang baru mempunyai perhatian untuk melakukan suatu gerakan. Pada tahap kesiapan, seseorang menunjukkan kesiapan mental dan fisik untuk melakukan suatu gerakan. Pada tingkatan meniru, seseorang sudah dapat meniru gerakan secara terbimbing. Pada tingkatan pembiasaan gerakan, seseorang melakukan gerakan secara mekanistik. Pada tingkatan mahir, seseorang sudah dapat melaksanakan kegiatan kompleks dan termodifikasi. Pada tingkatan adaptasi, seseorang sudah dapat melaksanakan gerakan alami yang diciptakan sendiri atas dasar gerakan yang diciptakan sebelumnya. Pada tingkatan terakhir, seseorang sudah dapat menciptakan sesuatu yang baru yang orisinal dan sukar ditiru oleh orang lain.

Dalam bidang pembelajaran, terdapat dua kelompok mata pelajaran yang mengandung aspek psikomotor. Kelompok pertama adalah kelompok mata pelajaran yang memerlukan penggunaan alat-alat praktikum (misalnya mata pelajaran fisika, kimia, biologi) atau alat-alat bengkel (misalnya mata pelajaran teknik mesin dan teknik elektro). Kelompok kedua adalah kelompok mata pelajaran yang menitikberatkan kepada gerak otot secara teratur, misalnya mata pelajaran olah raga dan mata pelajaran keterampilan, seperti keterampilan menjahit, memasak, dan sebagainya. Kelompok pertama terkait dengan adanya praktikum, sedangkan kelompok kedua terkait dengan adanya praktek lapangan.

INSTRUMEN RANAH PSIKOMOTOR

Menurut Luneta, dkk (Djemari Mardapi, dkk, 2002: 35), instrumen untuk mengukur aspek psikomotor yang berkaitan dengan penggunaan alat dapat berupa: (1) tes *paper and pencil*, (2) tes identifikasi, (3) tes simulasi, dan (4) tes unjuk kerja (*performance test*).

Perhatikanlah bahwa instrumen ranah psikomotor termasuk ke dalam kelompok tes, bukan non tes, sebab pada pengukuran ranah psikomotor terdapat langkah-langkah atau cara-cara yang benar dan ada langkah-langkah atau cara-cara yang kurang benar. Kadang-kadang tes di ranah psikomotor disebut sebagai tes perbuatan.

Tes *Paper and Pencil*

Pada tes *paper and pencil*, walaupun bentuk aktivitasnya seperti tes tertulis, namun sasarannya adalah kemampuan siswa dalam menampilkan karya, misalnya berupa desain alat, desain grafis, dan semacamnya.

Contoh 8.1

Berikut ini adalah contoh *paper and pencil test* yang meminta siswa untuk membuat desain rumah sederhana.

Gambarlah desain rumah sederhana dengan persyaratan sebagai berikut. Ukuran bangunan 15 m × 8 m, terdiri dari 1 kamar tamu. 2 kamar tidur utama, 1 kamar tidur pembantu, 1 dapur, 1 ruang keluarga. 1 carport yang dapat memuat 2 mobil.

Tes Identifikasi

Tes identifikasi ditujukan untuk mengukur kemampuan siswa dalam mengidentifikasi sesuatu hal. misalnya menemukan bagian yang tidak berfungsi pada suatu alat.

Tes Simulasi

Tes simulasi dipakai untuk memperagakan penampilan siswa dalam suatu simulasi, sehingga dengan simulasi dapat dinilai apakah seorang siswa telah menguasai keterampilan tertentu, misalnya keterampilan menyetir mobil.

TES UNJUK KERJA (*PERFORMANCE TEST*)

Tes unjuk kerja dilakukan dengan menggunakan alat yang sesungguhnya untuk mengetahui apakah siswa sudah terampil menggunakan alat tersebut atau belum. Termasuk dalam hal ini, misalnya, menyuruh siswa SMK untuk mengelas, memasak, dan semacamnya. Tes semacam ini sering disebut ***performance test.***

AERA, APA, dan NCME (1999) mendefinisikan tes unjuk kerja[1] ***"require students to complete a process or produce a product in a context that closely resembles real-life situation"***. Berarti tes unjuk kerja meminta peserta tes untuk mengerjakan suatu proses atau menciptakan produk yang mana proses dan produk tersebut haruslah seperti yang terjadi atau mendekati dengan situasi kehidupan nyata.

Reynolds, Livingstone, dan Willson (2010:255) memberikan petunjuk kapan seseorang memilih tes unjuk kerja sebagai berikut.

1. ***Select performance assessment tasks that provide the most direct assessment of the educational objektive you want to measure.***
2. ***Select performance assessment tasks that maximize your ability to generalize the results of the assessment,***

---

[1] Ada yang menyebut tes unjuk kerja (*performance test*) sebagai tes otentik (*authentic assessment*) atau tes alternatif (*alternative test*)

3. *Select performance assessment tasks that reflect essential skills.*
4. *Select performance assessment tasks that encompass more than one learning objective.*
5. *Select performance assessment tasks that focus your evaluation on the processes and/or products you are most interested in.*
6. *Select performance assessment tasks that provide degree of realism.*
7. *Select performance assessment tasks that measure skills that are "teachable".*
8. *Select performance assessment tasks that are fair to all students.*
9. *Select performance assessment tasks can be assessed given the time and resources available.*
10. *Select performance assessment tasks that can be score in a reliable manner.*
11. *Select performance assessment tasks that reflect educational objectives that cannot be measured using more traditional measures.*

PROSEDUR PENGUKURAN RANAH PSIKOMOTOR

Untuk melakukan pengukuran hasil belajar pada aspek psikomotor, ada dua hal yang perlu diperhatikan, yaitu: (1) pembuatan soal atau perintah untuk melakukan sesuatu, dan (2) pembuatan instrumen untuk mengamati jawaban atau respons siswa. Soal atau perintah untuk hasil belajar aspek psikomotor dapat berupa soal, lembar kerja, lembar tugas, perintah kerja, atau lembar eksperimen. Di sisi lain, instrumen untuk mengamati jawaban atau respons siswa dapat berupa lembar observasi atau lembar penilaian. Lembar observasi atau lembar penilaian tersebut dapat berupa daftar cek (***check list***) yang biasanya merupakan *numerical rating scale* atau *descriptive rating scale*.

Daftar cek berisi seperangkat butir soal yang mencerminkan rangkaian tindakan atau perbuatan yang harus ditampilkan oleh peserta ujian dan terdiri dari indikator-indikator atau keterampilan-keterampilan dari aspek yang akan diukur. Dengan melakukan pengamatan terhadap subjek yang dinilai, penilai dapat membubuhkan tanda cek pada tempat yang disediakan. Tempat yang disediakan untuk tanda cek, dapat berupa pernyatan "ya" dan "tidak". dapat pula berupa rating, misalnya dalam skala 5 (misalnya sangat tepat, tepat, agak tepat, tidak tepat, sangat tidak tepat), skala 4 (misalnya sangat baik, baik, cukup, kurang), atau skala 3 (misalnya baik, cukup, kurang).

Contoh 8.2

Berikut ini dicontohkan daftar cek untuk melakukan pengukuran mengenai kemampuan praktik olah raga bola volley (diambil dari Permendikbud Nomor 104 Tahun 2014)

Nama Peserta	Keterampilan yang Dinilai																			
	Cara Service				Cara Passing Atas				Cara Passing Bawah				Cara Smash				Cara Membendung (Blok)			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Anton																				
Bertha																				
Charles																				
Dono																				
Dst																				

Kategori Penilaian: 4 = sangat baik
3 = baik
2 = cukup
1 = kurang

Pada Contoh 8.2, daftar ceknya dipakai untuk keseluruhan peserta tes. Kadang-kadang dibuat individual per masing-masing peserta tes.

Contoh 8.3

Berikut ini adalah contoh daftar cek untuk mengukur keterampilan bermain bola volley dalam *rating* skala lima yang bersifat indivitual.

Nama Siswa:

No	Indikator	Jawaban				
		1	2	3	4	5
1	Cara melakukan service					
2	Cara melakukan passing atas					
3	Cara melakukan passing bawah					
4	Cara melakukan smash					
5	Cara melakukan pembendungan (blocking)					

Keterangan: 5 = sangat tepat
4 = tepat
3 = agak tepat
2 = tidak tepat
1 = sangat tidak tepat

Pada Contoh 8.2 dan Contoh 8.3, penilai atau pengamat diminta membubuhkan tanda cek (√) pada tempat yang disediakan.

Salah satu kesulitan penilaian ranah psikomotor adalah pengamat harus mengamati peserta tes satu-per-satu, tidak bersamaan. Dengan demikian, jika terdapat 40 siswa dan misalnya hanya tersedia satu pengamat, maka pengamat tersebut harus mengamati satu-per-satu siswa sebanyak 40 kali. Jika setiap peserta tes memerlukan waktu 10 menit, maka untuk menguji 40 peserta tes diperlukan waktu hampir 7 jam. (Bandingkan dengan pengujian ranah kognitif yang dapat dilaksanakan secara serentak, sehingga dalam waktu 2 jam dapat dinilai ribuan peserta tes).

Lembar penilaian untuk penilaian ranah psikomotor dapat berisi sekumpulan indikator atau keterampilan aspek yang diukur dan tempat untuk memberikan skor terhadap indikator atau keterampilan yang diukur tanpa skala tertentu.

Contoh 8.4

Berikut adalah lembar penilaian untuk masing-masing siswa mengenai cara siswa bermain bola volley dengan cara memberi skor pada tempat yang disediakan. (Bandingkan dengan Contoh 8.3)

Nama Siswa :

No	Indikator	Skor*
1	Cara melakukan service	
2	Cara melakukan passing atas	
3	Cara melakukan passing bawah	
4	Cara melakukan smash	
5	Cara melakukan pembendungan (blocking)	

Keterangan: Skor diisi dengan bilangan 1 sampai dengan 5 dengan kualifikasi sebagai berikut.
5 = sangat tepat
4 = tepat
3 = agak tepat
2 = tidak tepat
1 = sangat tidak tepat

Pada contoh terakhir, penilai diminta untuk memberikan skor terhadap keterampilan peserta ujian yang diamati. Pada umumnya, orang lebih menyukai daftar cek seperti pada Contoh 8.2 daripada daftar isian seperti pada Contoh 8.3.

PENGEMBANGAN INSTRUMEN RANAH PSIKOMOTOR

Pada dasarnya proses pengembangan instrumen untuk ranah psikomotor sama seperti pada pengembangan instrumen ranah kognitif, seperti yang dibicarakan pada Bab IV.

PENSKORAN INSTRUMEN RANAH PSIKOMOTOR

Pada umumnya nilai untuk ranah psikomotor diwujudkan secara kuantitatif seperti pada nilai ranah kognitif.

Hal pertama yang harus diperhatikan adalah apakah ada pembobotan pada keterampilan yang dinilai. Misalnya apakah kelima keterampilan bermain bola volley di atas mempunyai bobot yang sama. Jika mempunyai bobot yang sama, maka penilai tinggal menjumlah skor dari masing-masing butir indikator. Jika tidak mempunyai bobot yang sama, maka diperlukan perhitungan yang lebih rumit dengan melakukan penghitungan rataan terbobot.

PENAFSIRAN HASIL PENGUKURAN RANAH PSIKOMOTOR

Setelah diperoleh skor untuk masing-masing indikator atau keterampilan yang diujikan, nilai untuk setiap peserta uji dapat diperoleh dengan menjumlah skor untuk seluruh indikator dibagi dengan skor maksimal yang mungkin dicapai.

Contoh 8.5

Misalnya skor Amir untuk bermain bola volley adalah sebagai berikut.

No	Indikator	Jawaban				
		1	2	3	4	5
1	Cara melakukan service		√			
2	Cara melakukan passing atas			√		
3	Cara melakukan passing bawah				√	
4	Cara melakukan smash				√	
5	Cara melakukan pembendungan (blocking)			√		

Berdasarkan lembar tersebut dapat dilihat bahwa skor Amir adalah 2 + 3 + 4 + 4 + 3 = 16. Pada hal skor maksimal yang mungkin dicapai adalah 25. Dengan demikian, nilai Amir pada permainan bola volley adalah $\frac{16}{25} = 64$.

Jika batas tuntas untuk permainan bola volley adalah 75, maka Amir belum tuntas, dan harus melakukan remedi untuk bermain bola volley.

BAHAN DISKUSI

1. Bandingkanlah cara penilaian pada aspek kognitif. pada aspek afektif. dan pada aspek psikomotor. Manakah yang menurut Anda paling praktis? Mengapa?

2. Tulislah keunggulan dan kelemahan tes unjuk kerja (*performance test*)

3. Misalnya Anda diminta untuk menilai cara siswa menggunakan mesin jahit dalam menjahit (misalnya pada SMK Tata Busana). Buatlah daftar cek untuk keperluan tersebut.

4. Misalnya Anda diminta untuk menilai cara *start* pada lari cepat 100 meter (pada mata pelajaran Olah Raga). Buatlah daftar cek untuk keperluan tersebut.

5. Salah satu kelemahan penilaian unjuk kerja adalah diperlukan waktu yang cukup banyak. Bagaimana cara untuk mengatasi kelemahan tersebut?

****

BAB IX
PENILAIAN BERBASIS KELAS, PENILAIAN UNTUK PEMBELAJARAN, DAN PENILAIAN OTENTIK

PENDAHULUAN

Seperti disebutkan pada Bab I, Johnson & Johnson (2002) menggolongkan penilaian ke dalam tiga jenis, yaitu: penilaian diagnostik, penilaian formatif, dan penilaian sumatif. Dengan penilaian diagnostik, para pendidik diharapkan dapat mengetahui kesalahan dan/atau miskonsepsi yang terjadi pada peserta didik. Penilaian formatif adalah penilaian yang bertujuan untuk memberikan balikan kepada peserta didik terkait dengan kemajuan yang telah ia capai dan untuk memberikan balikan kepada pendidik terkait dengan perkembangan proses pembelajaran yang dirancangnya. Penilaian sumatif dilakukan dengan tujuan untuk menentukan kedudukan peserta didik terkait dengan hasil pembelajaran yang telah diperolehnya. Penilaian sumatif biasanya berbentuk ujian semester atau ujian akhir satuan pendidikan. Penilaian yang didefinisikan oleh Popham (1995: 5), seperti yang ditulis di Bab I, lebih mengarah ke definisi penilaian sumatif daripada definisi jenis penilaian yang lain.

Penggolongan lain penilaian adalah membagi penilaian ke dalam dua tipe, yaitu penilaian internal dan penilaian eksternal. Penilaian internal adalah penilaian yang dilakukan oleh pendidik kepada peserta didiknya sendiri, sedangkan penilaian eksternal adalah penilaian yang dilakukan oleh lembaga di luar lembaga pendidik berdasarkan kepada pedoman yang telah disepakati. Ujian nasional adalah salah satu contoh penilaian eksternal yang dilakukan oleh Pemerintah. Di beberapa negara ada kebiasaan melakukan *benchmarking*, suatu penilaian eksternal yang mempunyai tujuan untuk melakukan penilaian terhadap suatu hal.

Pada pelaksanaan sehari-hari di lapangan, penilaian kadang diartikan berbeda tergantung kepada konteks dan siapa yang mengartikannya. Seperti yang dikatakan oleh Garfield (1994), kebanyakan pendidik (guru dan dosen) mengartikan penilaian "*in terms of testing and grading: scoring quizzes and exams and assigning course grade to students*". Jadi, penilaian diartikan dalam arti sempit yaitu sekedar pemberian tes dan pemberian nilai, kegiatan penilaian hanyalah kegiatan melakukan skoring pada kuis dan ujian untuk memberikan nilai kepada siswa (mahasiswa). Lebih lanjut Garfield mengatakan bahwa kebanyakan pendidik (guru dan dosen) "*use assessment as a way to inform students about how well they are doing or how well they did in the courses we teach*". Mereka menggunakan penilaian sebagai suatu cara untuk memberitahukan kepada siswa seberapa baik yang telah mereka kerjakan dan/atau memberitahukan kepada siswa seberapa baik mereka menguasai mata pelajaran atau mata kuliah yang telah diajarkan oleh guru atau dosennya Kalau ini yang terjadi, maka ini berarti bahwa penilaian hanya dipandang sebagai penilaian sumatif.

Memandang penilaian hanya sebagai penilaian sumatif memberikan dampak yang tidak menguntungkan. Dampak-dampak tersebut antara lain:

(1) memisahkan kegiatan penilaian dengan kegiatan pembelajaran, yang hal ini tampak jelas ketika para pendidik membuat RPP (rencana pelaksanaan pembelajaran), di mana pendidik menempatkan kegiatan penilaian setelah kegiatan pembelajaran selesai (*assessments take place after instructions*),

(2) tujuan utama penilaian hanya untuk membuat rangking, untuk membedakan siswa yang pandai dan siswa yang tidak pandai, untuk membedakan siswa yang lulus dan siswa yang tidak lulus, untuk membedakan siswa mana yang berhak mendapat beasiswa dan yang tidak, dan tindakan-tindakan diskriminatif lainnya,

(3) penilaian sering dipakai untuk menghukum peserta didik, misalnya dengan memberikan nilai jelek pada mata pelajaran yang diampu oleh pendidik.

(4) penilaian tidak membantu peserta didik yang mempunyai kesulitan belajar, sehingga tidak dapat menciptakan *equity* di dalam pendidikan.

Sejak tahun duaribuan, di kalangan praktisi pendidikan terjadi kegundahan akibat adanya penyempitan pengertian mengenai penilaian tersebut. Memandang penilaian hanya sebagai penilaian sumatif tidaklah menguntungkan kepada peningkatan kualitas pembelajaran. Diperlukan pandangan baru terhadap penilaian agar penilaian merupakan kegiatan yang menyatu dengan kegiatan pembelajaran yang pada ujungnya dapat meningkatkan kualitas pembelajaran. Dari sini munculah berbagai nama penilaian yang membedakan dengan penilaian yang sekarang ini banyak dipahami orang.

Nama-nama baru penilaian itu, yang lebih menekankan kepada penilaian formatif, misalnya penilaian berbasis kelas (*classroom assesment*) dan penilaian untuk pembelajaran (*assessment for learning*). Di sisi lain, pandangan yang menitikberatkan penilaian hanya sebagai penilaian sumatif dan yang hanya berupa *paper and pencil test* sering disebut orang sebagai penilaian tradisional (*traditional assesment*).

Pada praktiknya, memang penilaian tradisional tersebut masih tetap diperlukan, tetapi pelaksanaannya harus dibarengi dengan penilaian alternatif yang dapat meningkatkan kualitas pembelajaran, yang lebih bersifat formatif, yang dapat membantu siswa yang berkesulitan belajar untuk memperbaiki kesalahannya.

PENILAIAN BERBASIS KELAS (*CLASSROOM ASSESMENT*)

Ada berbagai definisi penilaian berbasis kelas, yang antara satu definisi dengan definisi lainnya kadang saling bertolak belakang.

Badan Standar Nasional Pendidikan (BSNP) (Nuning Hidayah Sunani, 2010: 65) menyatakan bahwa penilaian berbasis kelas merupakan "suatu kegiatan yang dilakukan oleh guru berupa pengumpulan informasi selama pembelajaran berlangsung melalui prosedur, alat penilaian, dan berbagai teknik yang sesuai dengan kompetensi yang akan dinilai". Jika definisi ini yang dipakai, maka semua kegiatan penilaian yang dilakukan oleh guru di kelas disebut penilaian berbasis kelas. Dengan demikian, maka berbagai bentuk penilaian (penilaian yang manapun juga) merupakan penilaian berbasis kelas, jika dilakukan oleh guru di dalam kelas. Menurut definisi ini, penilaian disebut penilaian yang tidak berbasis kelas apabila tidak dilakukan oleh guru di kelas. Ujian nasional, misalnya, bukanlah penilaian berbasis kelas, tetapi ujian pilihan ganda yang dilakukan oleh guru di kelas merupakan penilaian berbasis kelas.

Di sisi lain, Angelo (Nuning Hidayah Sunani, 2010: 64) menyatakan bahwa *"classroom assessment consist of small scale assessment conducted continuously in college classrooms to determine what students are learning in that class"*.

Lebih lanjut, senada dengan Angelo, dikatakan bahwa:

> *Classroom assessment is both a teaching approach and a set of techniques. The approach is that the more you know about what and how students are learning, the better you can plan learning activities to structure teaching. The techniques are mostly simple, non-graded, anonymous, in-class activities that give both you and your students useful feedback on the teaching-learning process* (http://ntlf.com/ html/lib/bib/ assess.htm, diambil 2 Mei 2010).

Definsi Angelo ini menunjukkan bahwa penilaian berbasis kelas adalah penilaian formatif. Yang terpenting dari penilaian berbasis kelas adalah adanya umpan balik kepada peserta didik. Umpan balik tersebut juga dapat pula mengena kepada guru manakala guru menggunakan hasil penilaian berbasis kelas unuk memperbaiki proses pembelajarannya. Mengacu kepada definisi Angelo, maka wujud dari penilaian berbasis kelas adalah simpel, mungkin berupa ujian singkat di kelas yang bisa diselesaikan dengan cepat, kemudian hasil ujian para peserta didik diperiksa dan diberi umpan balik sekiranya hasil ujian para peserta didik belum memenuhi kriteria yang diharapkan.

Penjelasan mengenai penilaian berbasis kelas seperti yang didefinisikan oleh Angelo di atas menekankan pentingnya *feedback* dalam pembelajaran sehari-hari dalam rangka memperbaiki kesalahan-kesalahan yang diperbuat oleh siswa. Wujud dari penilaian berbasis kelas adalah *simple* dan *non-grade* serta berlangsung secara terus menerus dalam suatu proses pembelajaran. Penilaian berbasis kelas dapat lisan (*oral*) maupun tertulis (*written*).

Bagi guru, adanya penilaian berbasis kelas seperti yang didefinisikan oleh Angelo memberi keuntungan, antara lain: (1) memberikan umpan balik mengenai proses pembelajaran dan dengan segera dapat memperbaikinya manakala ada hambatan yang muncul, (2) memberi informasi berharga mengenai cara belajar peserta didiknya, (3) mendorong pemahaman bahwa mengajar adalah proses formatif yang melibatkan umpan balik (feed back) secara terus menerus.

PENILAIAN UNTUK PEMBELAJARAN (*ASSESSMENT FOR LEARNING*)

Berdasarkan riset kecil-kecilan yang telah dilakukan oleh penulis, melalui berbagai wawancara dengan para guru, diperoleh temuan bahwa para guru telah merasa memberikan penilaian kepada siswanya dengan baik, baik penilaian formatif (yang disebut dengan ulangan harian) maupun penilaian sumatif (yang disebut ulangan akhir semester).

Terkait dengan penilaian formatif (yang oleh beberapa guru disebut ulangan harian), dapat disampaikan temuan berikut.

a. Setiap guru telah melaksanakan ulangan harian setiap satu kompetensi dasar (KD) dilaksanakan;

b. Setiap guru telah memeriksa ulangan harian tersebut, memberikan skor pada lembar pekerjaan siswa, dan membagikan kembali kepada siswa;

c. Skor yang diberikan kepada siswa lebih berfungsi sebagai bagian dari pemberian nilai kepada siswa, karena ikut dihitung untuk menentukan nilai akhir rapor, bukan berfungsi sebagai balikan;

d. Rentang waktu antara pelaksanaan ulangan harian dengan pengembalian hasil pemeriksaan lembar pekerjaan bervariasi. paling cepat seminggu;

e. Hampir sebagian besar guru tidak memberikan balikan kepada siswa di dalam lembar pekerjaan siswa. misalnya memberitahukan bahwa siswa yang bersangkutan melakukan kesalahan dan bagaimana memperbaiki kesalahan tersebut;

f. Hampir sebagian besar guru tidak memberikan pujian kepada siswa, apabila ada siswa yang mengerjakan dengan baik;

g. Bagi siswa yang mendapatkan skor jelek pada KD tersebut, maka diberikan ulangan harian kembali, yang oleh para guru kegiatan memberikan ulangan kembali tersebut memberikan remediasi.

Berdasarkan temuan penelitian tersebut. menurut hemat penulis, dapat disimpulkan bahwa para guru belum melaksanakan penilaian formatif dengan benar, karena fungsi penilaian formatif sebagai wahana untuk memberikan balikan (*feed-back*) kepada siswa secepat mungkin belum tampak benar pada pelaksanaan penilaian yang dilakukan oleh para guru.

Kecuali disebutkan di atas, di kalangan praktisi pendidikan di Indonesia, perhatian lebih ditekankan kepada bagaimana mengkonstruksi penilaian sumatif (yang juga disebut *assessment of learning*, AoL) yang baik, misalnya pada penyusunan soal-soal ujian nasional dan pada ujian masuk perguruan tinggi. Kepada para guru pun banyak dilatihkan bagaimana mengkonstruksi AoL yang baik, misalnya pada penyusunan soal-soal untuk ulangan umum bersama. Kuliah-kuliah penilaian pembelajaran di perguruan tinggi juga lebih dititikberatkan kepada hal ihwal mengenai penilaian sumatif dibandingkan dengan penilaian formatif.

Di sisi lain, dewasa ini di dunia penilaian, telah lama dikembangkan salah satu jenis penilaian yang disebut penilaian untuk pembelajaran (*assessment for learning*, untuk selanjutnya disingkat AfL). AfL ini pada dasarnya adalah penilaian formatif. Diberi nama AfL dengan tujuan untuk menekankan bahwa penilaian yang dilakukan adalah penilaian untuk perbaikan pembelajaran. bukan penilaian untuk melihat seberapa banyak pengetahuan yang telah dikuasai oleh siswa.

Dalam salah satu makalahnya, Young (2005) mengatakan bahwa AfL, jika digunakan secara efektif, dapat meningkatkan prestasi siswa. Hal yang sama dikemukakan oleh Stiggins & Chappuis (2006) bahwa AfL dapat meningkatkan kesuksesan siswa. Di Inggris, AfL sudah diterapkan sejak lama dan terbukti telah dapat meningkatkan kemampuan matematika siswa.

Assessment Reform Group di Inggris yang disponsori oleh *British Educational Research Association* telah melakukan riset mendalam mengenai pelaksanaan AfL di Inggris sejak beberapa lama. Mereka

mengklaim bahwa AfL dapat meningkatkan kemampuan siswa dalam berbagai mata pelajaran, seperti yang dikemukan oleh Young (2005) dan Stiggins & Chappuis (2006).

Assessment for learning (AfL) didefinisikan sebagai *using evidence and feedback to identify where students are in their learning, what they need to do next, and how best to achieve this* (www.geography.org.uk). Dengan kata lain, AfL adalah *the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go, and how best to get there*. Berdasarkan hal-hal tersebut dapat dikatakan bahwa AfL adalah proses untuk mencari dan menginterpretasikan bukti-bukti yang ada untuk digunakan bagi siswa dan guru untuk menentukan pada posisi mana siswa-siswa telah belajar, apa yang harus dikerjakan kemudian, dan bagaimana cara terbaik untuk mencapai tujuan yang diinginkan.

AfL dikembangkan berdasar kepada pemikiran bahwa kemampuan siswa dapat meningkat secara optimal, jika mereka mengerti tujuan pembelajaran, mengetahui posisi mereka dalam kaitannya dengan tujuan pembelajaran, dan mengerti cara mencapai tujuan pembelajaran tersebut.

Ada 10 prinsip dalam AfL, yaitu:

(1) AfL merupakan bagian dari perencanaan pembelajaran yang efektif (AfL *should be part of effective planning of teaching and learning*),
(2) AfL harus menfokuskan kepada bagaimana siswa belajar (AfL *should focus on how students learn*),
(3) AfL harus merupakan pusat dari praktik pembelajaran di kelas (AfL *should be recognized as central to classroom practice*),
(4) AfL merupakan kunci keterampilan profesional guru (AfL *should be regarded as a key professional skill for teachers*),
(5) AfL harus sensitif dan konstruktif, sebab setiap asesman selalu mempunyai dampak emosional kepada siswa (AfL *should be sensitive and constructive because any assessment has an emotional impact*),
(6) AfL harus memperhatikan pentingnya motivasi siswa (AfL *should take account of the importance of learner motivation*),
(7) AfL harus mengutamakan komitmen atas tujuan pembelajaran dan pemahaman mengenai kriteria yang harus dinilai (AfL *should promote commitment to learning goals and a shared understanding of the criteria by which they are assessed*),
(8) Pada AfL, siswa harus mendapatkan petunjuk konstruktif bagaimana siswa harus memperbaiki diri (*learner should receive constructive guidance about how to improve*),

(9) AfL harus dapat mengembangkan kapasitas siswa untuk dapat menilai dirinya sendiri (AfL *should develops learners' capacity for self-assessment so that they can become reflective and self managing*), dan

(10) AfL harus mengenal rentang kemampuan siswa (AfL *should recognise the full range of achievement of all learners*).

Ada empat karakteristik kunci yang harus dipahami oleh guru dalam melaksanakan AfL, yaitu:

(1) digunakannya teknik bertanya yang efektif (*using effective questioning techniques*),

(2) digunakannya strategi pemberian balikan (*using feedback strategies*)

(3) adanya pengertian bersama mengenai tujuan pembelajaran (*sharing learning goals*).

(4) dilakukannya penilaian antar teman dan penilaian diri (*peer and self-assesment*).

Untuk mewujudkan AfL yang efektif, hal-hal berikut harus dilakukan oleh guru:

(1) menekankan adanya interaksi antara pembelajaran dan penilaian yang dapat meningkatkan kualitas pembelajaran (*emphases the interactions between learning and manageable assessment strategies that promote learning*),

(2) menyatakan secara jelas tujuan pembelajaran (*clearly expresses for the student and teacher the goals of the learning activity*),

(3) menyatakan pandangan belajar bahwa penilaian dapat membantu siswa belajar lebih baik, bukan sekedar memperoleh nilai yang baik (*reflects a view of learning in which assessment helps students learn better, rather than just achieve a better mark*),

(4) memberikan arahan kepada siswa dengan memberikan balikan kepada mereka (*provides ways for students to use feedback from assessment*),

(5) membantu siswa untuk bertanggung jawab mengenai kemajuan belajarnya sendiri (*helps students take responsibility for their own learning*).

(6) berlaku untuk seluruh siswa (*is inclusive of all learners*).

Di sisi lain, Clarke (2005: 1-2) mengatakan bahwa pelaksanaan AfL (yang oleh Clarke disebut penilaian formatif) harus mengikuti strategi berikut:

(1) menyatakan dengan jelas tujuan pembelajaran dan kriteria sukses pada perencanaan pembelajaran sebagai kerangka dasar untuk AfL (*clarifying learning objectives and success criteria at the planning stage, as framework for formative assessment processes*),

(2) berbagi tujuan pembelajaran dan kriteria sukses dengan siswa (*sharing learning objectives and success criteria with students, both long term and for individual lessons*),
(3) menggunakan teknik bertanya yang tepat dan efektif untuk mengembangkan pembelajaran, bukan untuk mengukur kemampuan siswa (*appropriate and effective questioning which develops the learning rather than attempts to measure it*),
(4) memusatkan kepada pemberian balikan, baik secara lisan maupun tertulis (*focusing oral and written feedback, whether from teacher or student, around the development of learning objectives and meeting of targets*),
(5) menata target sedemikian hingga pencapaian kemampuan siswa berdasarkan kepada kemampuan sebelumnya (*organising targets so that students' achievement is based on previous achievement as well as aiming for the next step*),
(6) melibatkan penilaian diri dan penilaian antar-teman (*involving students in self- and peer evaluation*), dan
(7) memberikan pemahaman bahwa setiap siswa dapat belajar dan berkembang dengan baik (*raising students' self-efficacy and holding a belief that all students have potential to learn and achieve*).

Seperti diuraikan, inti dari AfL adalah pemberian balikan kepada siswa secepat mungkin terhadap kesalahan-kesalahan yang dilakukan oleh siswa. Wujudnya dapat bermacam-macam. Termasuk *peer assessment* yaitu penilaian antarteman.

PENERAPAN *ASSESSMENT FOR LEARNING* (AfL) DI KELAS

Berikut ini adalah **suatu contoh** model AfL yang dikembangkan penulis bersama tim bekerjasama dengan Musyawarah Guru Mata Pelajaran (MGMP) Kota Surakarta. Tentu saja, di sana-sini, model yang telah dikembangkan tersebut dapat disempurnakan dan/atau dimodifikasi untuk memenuhi asas kepraktisan (kemudahan penggunaan). Model tersebut dikembangkan melalui *Research and Development* yang dibiayai oleh DIPA DIKTI melalui DIPA UNS pada skema penelitian potensi pendidikan dengan nomor kontrak 0162.0/023-04.2/XIII/2008, tanggal 31 Desember 2008.

Berdasarkan pengertian dan prinsip-prinsip AfL yang disampaikan di depan, model AfL yang telah dikembangkan oleh penulis dan teman-teman pada pembelajaran matematika di SLTP di Kota Surakarta, mengikuti strategi dan implementasi seperti pada Tabel 9.1.

Tabel 9.1. Implementasi Strategi AfL dari Clarke

No	Strategi AfL dari Clarke (2005)	Implementasi
1	*Clarifying learning objectives and success criteria at the planning stage, as framework for formative assessment processes*	Memformulasikan tujuan pembelajaran dan kriteria sukses sebelum pembelajaran berlangsung. Tujuan pembelajaran dan kriteria sukses mengacu kepada RPP
2	*Sharing learning objectives and success criteria with students, both long term and for individual lessons*	Memberitahukan tujuan pembelajaran dan kriteria sukses kepada siswa di setiap awal pembelajaran dan menulisnya di papan tulis, sehingga selama pembelajaran berlangsung guru dan siswa dapat memfokuskan kepada tujuan pembelajaran dan kriteria sukses tersebut
3	*Appropriate and effective questioning which develops the learning rather than attempts to measure it*	Menggunakan tujuan pembelajaran dan kriteria sukses sebagai dasar untuk memberikan pertanyaan (*questioning*) dan balikkan (*feed-back*) selama pembelajaran berlangsung
4	*Focusing oral and written feedback, whether from teacher or student, around the development of learning objectives and meeting of targets*	Memeriksa hasil pekerjaan siswa sesegera mungkin. Memberikan balikan konstruktif dan memotivasi kepada siswa pada lembar pekerjaan siswa
5	*Organising targets so that students' achievement is based on previous achievement as well as aiming for the next step*	Menetapkan tujuan pembelajaran dan kriteria sukses berdasarkan pencapaian kemampuan siswa sebelumnya
6	*Involving students in self- and peer evaluation*	Memberi kesempatan kepada siswa untuk mendiskusikan dengan teman-temannya jawaban atas soal-soal yang diberikan, mengemukakan perasaannya mengenai pembelajaran yang sedang berlangsung dan mengemukakan kesulitan-kesulitan yang ditemui selama pembelajaran
7	*Raising students' self-efficacy and holding a belief that all students have potential to learn and achieve*	Selama pembelajaran berlangsung, guru memberikan semangat dan membangun kepercayaan diri siswa bahwa setiap siswa dapat belajar matematika dengan baik. Memberikan soal-soal mulai dari soal yang mudah, sehingga setiap siswa merasa dapat mengerjakan soal dengan benar. Memberikan balikan yang konstruktif.

Berdasarkan strategi dan implementasi prinsip-prinsip AfL di atas, maka perencanaan dan pengorganisasian pembelajaran yang digunakan dalam model AfL disusun sebagai berikut.

a. Perencanaan Pembelajaran

Untuk dapat mewujudkan AfL yang menyatu dengan proses pembelajaran, guru mempersiapkan hal-hal berikut.

1. *Tujuan pembelajaran*. Tujuan pembelajaran disusun dalam kalimat yang dapat dipahami oleh siswa. Supaya siswa selalu mengingat tujuan pembelajaran yang harus dicapai, tujuan pembelajaran ditulis di papan tulis dan tidak dihapus selama pembelajaran berlangsung.
2. *Kriteria sukses*. Guru menetapkan kriteria sukses sebagai kriteria bahwa siswa telah berhasil mencapai tujuan yang dirumuskan, misalnya siswa dikatakan sukses apabila dapat mengerjakan soal-soal yang ditentukan. Seperti halnya tujuan pembelajaran, soal-soal yang diharapkan dapat diselesaikan oleh siswa ditulis di papan tulis dan tidak dihapus selama pembelajaran berlangsung.
3. *Soal-soal latihan*. Guru menyiapkan tiga jenis soal sebagai latihan, merupakan soal uraian (*essay*), yang disebut soal tahap I, soal tahap II, dan soal tahap III, masing-masing minimal sebuah soal. Jika dirasa soal tahap I dan tahap II sudah cukup, soal tahap III tidak perlu diberikan. Pemberian soal tahap I dan tahap II ini diadopsi dari model AfL yang dikembangkan oleh Mansyur (2009) yang juga merupakan modifikasi dari *two-stage tasks* dari de Lange (1999). Perbedaan tiga jenis soal tersebut tampak pada Tabel 9.2.

Tabel 9.2. Perbedaan Soal Tahap I, Tahap II, dan Tahap III

No	Aspek	Soal Tahap I	Soal Tahap II	Soal Tahap III
1	Tingkat kesulitan	Mudah	mudah atau sedang	sedang atau sukar
2	Lama pengerjaan	10 – 15 menit	20 – 30 menit	1 – 2 jam
3	Pengerjaan	di kelas	di rumah	di rumah
4	Waktu penyerahan kepada guru	diserahkan di kelas, langsung setelah selesai mengerjakan	diserahkan kepada guru sehari sebelum pembelajaran berikutnya	tidak diserahkan kepada guru

Tabel 9.2. Perbedaan Soal Tahap I, Tahap II. dan Tahap III (lanjutan)

No	Aspek	Soal Tahap I	Soal Tahap II	Soal Tahap III
4	Waktu penyerahan kepada guru	diserahkan di kelas, langsung setelah selesai mengerjakan	diserahkan kepada guru sehari sebelum pembelajaran berikutnya	tidak diserahkan kepada guru
5	Waktu pemeriksaan pekerjaan siswa	diperiksa langsung oleh guru di kelas (atau oleh tim) dan dikembalikan kepada siswa pada saat itu juga	diperiksa oleh guru tidak di kelas, dikembalikan kepada siswa pada pembelajaran berikutnya	tidak diperiksa oleh guru
6	umpan balik	diberikan umpan balik bagi yang melakukan kesalahan	diberikan umpan balik bagi yang melakukan kesalahan	didiskusikan di kelas sebagai wahana pemberian umpan balik
7	pemberian skor	diberi skor, tetapi skor tidak diberitahukan kepada siswa	diberi skor, tetapi skor tidak diberitahukan kepada siswa	tidak diberi skor
8	fungsi pemberian skor	untuk merekam kemajuan siswa, bukan sebagai bagian dari pemberian nilai kepada siswa	untuk merekam kemajuan siswa, bukan sebagai bagian dari pemberian nilai kepada siswa	
9	pemberian umpan balik dan motivasi	diberi umpan balik dan diberi pujian untuk memberi motivasi	diberi umpan balik dan diberi pujian untuk memberi motivasi	
10	jenis pujian dan balikan	pada lembar pekerjaan siswa ditulis: ***excellent***: jika benar dikerjakan dengan sempurna ***good***: jika hampir benar ***perbaiki***: jika salah (tunjukkan bagaimana cara memperbaikinya)	Pada lembar pekerjaan siswa ditulis: ***excellent***: jika benar dikerjakan dengan sempurna ***good***: jika hampir benar ***perbaiki***: jika salah (tunjukkan bagaimana cara memperbaikinya)	

Perhatikan bahwa soal tahap III berfungsi sebagai pekerjaan rumah seperti yang biasanya diberikan oleh guru kepada siswanya. Jawaban soal-soal tahap III ini didiskusikan di kelas, tetapi pada model ini tidak ada kewajiban bagi guru untuk memeriksa dan memberi balikan secara tertulis kepada masing-masing siswa.

Karena soal tahap I dan soal tahap II merupakan soal-soal untuk penilaian formatif, maka sebenarnya tidak perlu diberi skor. Pemberian skor diperlukan jika guru ingin mengetahui perkembangan kemajuan siswanya. Pada praktik pembelajaran sehari-hari, disarankan soal tahap I dan soal tahap II tidak perlu diberikan skor, tetapi diberi *feedback*. Seperti dikatakan de Lange (1999), "*feedback can be immediate and very differentiated in the sense that the feedback can be direct (giving the student information about what is wrong and why and giving a suggestion for correction)*". Jadi, feedback harus diberikan segera mungkin dan bersifat individual, berisi informasi mengenai kesalahan yang dilakukan dan saran bagaimana memperbaiki kesalahan tersebut.

Beberapa orang mengatakan bahwa pemberian skor kepada kertas pekerjaan siswa adalah wujud dari suatu *feedback*. Pendapat ini tidak sepenuhnya benar, sebab menurut de Lange (1999) "*a score on a test is encoded information, whereas feedback is information that provides the performer with direct, usable insights into current performance and is based on tangible differences between current performance and hoped-for performance*".

Skor-skor untuk soal tahap I dan soal tahap II tersebut, jika diberikan, tidak dipakai sebagai pertimbangan pemberian nilai akhir (nilai rapor). Jika guru menginginkan ada skor-skor lain di samping skor ujian akhir semester untuk menentukan nilai rapor siswa, maka guru dapat memberikan ulangan harian yang sifatnya sebagai penilaian sumatif (atau penilaian sub-sumatif). Penilaian sub-sumatif ini diberikan setelah beberapa kompetensi dasar (KD) selesai diajarkan.

b. Pengorganisasian Pembelajaran

Untuk melaksanakan pembelajaran dengan AfL ini, pembelajaran matematika di kelas diatur sedemikian rupa sehingga satu satuan pembelajaran berlangsung selama 2 jam pelajaran (2×40 menit). Selama 2×40 menit, yang diharapkan melaksanakan pembelajaran untuk satu kompetensi dasar (KD) (atau bagian dari satu KD), pengorganisasian pembelajaran disusun seperti pada Tabel 9.3.

Tabel 9.3. Lama Waktu, Kegiatan Guru, dan Kegiatan Siswa dalam AfL

No	Kegiatan Guru	Kegiatan Siswa	Lama
1	a. Memberikan apersepsi dan motivasi. b. Menulis tujuan dan kriteria sukses di papan tulis. c. Menjelaskan tujuan pembelajaran dan kriteria sukses kepada siswa.	Memahami dengan baik tujuan pembelajaran dan kriteria sukses yang disampaikan guru	5 menit
2	Melaksanakan pembelajaran sesuai dengan RPP yang dibuat guru	Melaksanakan pengalaman belajar sesuai dengan RPP yang dibuat guru	40 – 45 menit
3	Memberikan soal tahap I	Mengerjakan soal tahap I di kelas	10 – 15 menit
4	Memeriksa jawaban siswa untuk soal tahap I dan memberikan balikan pada kertas jawaban siswa (oleh guru kelas maupun bersama-sama dengan guru lain dalam suatu *team teaching*, atau dengan cara lain) dan mengembalikan kertas jawaban kepada masing-masing siswa	Berdiskusi dengan teman-temannya mengenai jawaban soal tahap I. Beberapa siswa, misalnya 3 orang siswa, diminta menulis jawaban soal tahap I di papan tulis sebagai hasil diskusi dengan siswanya	10 – 15 menit
5	Memberikan balikan kepada siswa secara klasikal terhadap pengerjaan soal tahap I, secara lisan Menanggapi kesulitan-kesulitan yang dialami oleh siswa dalam mengerjakan soal	Mendengarkan dan mencatat balikan yang diberikan oleh guru. Mengemukakan kesulitan yang dialami oleh siswa dalam mengerjakan soal	5 – 10 menit
6	Memberikan soal tahap II dan tahap III	Mencatat soal tahap II dan tahap III (jika belum disediakan oleh guru)	5 menit

Catatan:

1) Pada pembelajaran berikutnya, sebelum memulai pembelajaran, guru membagi pekerjaan siswa untuk soal tahap II yang telah diberikan umpan balik
2) Guru perlu memberikan umpan balik terhadap pengerjaan soal tahap II secara lisan di kelas

3) Untuk memudahkan siswa melihat balikan guru, soal tahap I dan tahap II ditulis pada selembar kertas dan siswa diminta mengerjakan di kertas itu juga.

Perhatikanlah bahwa model pembelajaran dengan AfL ini, penilaian formatif (yang diwujudkan dalam soal-soal tahap I. dan II) benar-benar menyatu dengan proses pembelajaran. Tujuan utama penilaian tersebut adalah untuk memberikan balikan kepada siswa, jika siswa melakukan kesalahan, dan memberikan pujian, jika siswa mengerjakan soal dengan benar.

Setelah dilakukan uji coba pada skala terbatas (*preliminary field testing*) pada kelas VII, VIII, dan IX di SMP Negeri 14 dan SMP Muhammadiyah 1 Surakarta, model yang dikembangkan telah diimplementasikan pada skala luas (*main field testing*) pada kelas VII. VIII. dan IX di enam buah SMP di Kota Surakarta.

SMP-SMP yang terlibat dalam implementasi model adalah: SMP Negeri 14. SMP Negeri 18. SMP Negeri 19, SMP Negeri 20. SMP Muhammadiyah 1, dan SMP Kristen Kalam Kudus Surakarta. Implementasi model dilaksanakan pada bulan Oktober dan Nopember 2009 selama tiga kali pembelajaran.

Nama-nama guru yang terlibat pada langkah ini dapat dilihat pada Tabel 9.4.

Tabel 9.4. Nama SMP dan Guru yang Terlibat dalam Main Field Testing

No	Nama SMP	Nama Guru		
		Kelas VII	Kelas VIII	Kelas IX
1	SMP Negeri 14	Tri Purwandari, S.Pd.	Dra. Tri Unggul Suwarsi, M.Pd.	Yahya Irine, S.Pd.
2	SMP Negeri 18	Sri Wulandari, S.Pd.	Prih Sasonodadi, S.Pd.	Partini, S.Pd.
3	SMP Negeri 19	Tri Isnadi. S.Pd.	Mahanani Surjatiningsih, S.Pd.	Endang Sriningsih, S.Pd., M.Pd.
4	SMP Negeri 20	Diana Indriastuti KW, S.Pd., M.Pd.	Alip Tohar Mustakim, S.Si.	Murwaningsih. S.Pd.
5	SMP Muhammadiyah 1	Erwin Kurniati. S.Pd.	Agus Budi Hartono, S.Pd., M.Pd.	Hermawan Lastiyono, S.T., S.Pd.
6	SMP Kristen Kalam Kudus	Friesca Pra Utami Dewi. S.Pd.	Ivid Kristyana Savitri, S.T.	Evi Dayanti, S.Pd.

Keberjalanan model yang telah dikembangkan dilihat dari 4 aspek, yaitu dilihat dari: (a) guru yang menjalankan model tersebut, (b) siswa yang

dikenai pembelajaran, (c) pembandingan nilai siswa-siswa sebelum dikenai AfL dengan nilai-nilai mereka setelah dikenai AfL, dan (d) pembandingan nilai siswa-siswa yang dikenai AfL dengan siswa-siswa yang tidak dikenai AfL (dengan asumsi kelas yang dikenai AfL dan kelas yang tidak dikenai AfL sebanding kemampuannya).

Untuk melihat pembandingan pada (c) dan (d) hanya dilakukan secara *ex post facto*, tidak melalui penelitian eksperimental yang terkontrol dengan ketat. Para peneliti lain dapat melakukan pembandingan prestasi belajar antara kelas yang dikenai AfL dan kelas yang tidak dikenai AfL secara eksperimental yang terkontrol secara ketat untuk melihat efektivitas pembelajaran yang mengakomodasi AfL.

Pada implementasi tersebut, kepada para guru dipersilakan untuk mengajar dengan menggunakan Rencana Pelaksanaan Pembelajaran (RPP) yang telah mereka buat sebelumnya dengan mengadakan modifikasi dengan memasukkan model AfL pada pembelajarannya. Metode yang dipakai pada pembelajaran diserahkan sepenuhnya kepada para guru. Pokok bahasan yang digunakan pada implementasi juga diserahkan sepenuhnya kepada para guru. Dengan demikian, pembelajaran yang mengakomodasi model AfL ini bebas metode dan bebas pokok bahasan.

Pada pelaksanaan model tersebut, hampir semua guru mengatakan bahwa model AfL dapat dilaksanakan dengan cukup mudah. Aspek yang terasa membebani adalah pemberian pertanyaan yang bersifat untuk mendapatkan umpan balik, pemeriksaan pekerjaan siswa, dan pemberian balikan kepada siswa secara tertulis pada pekerjaan siswa.

Hal tersebut di atas menunjukkan bahwa keterampilan memberikan pertanyaan efektif untuk melihat hal-hal yang belum diketahui siswa bukan pekerjaan yang mudah dan perlu dikembangkan terus menerus. Pemeriksaan pekerjaan siswa dan pemberian balikan tertulis kepada siswa selama ini juga belum merupakan kebiasaan guru dalam mengajar. Pada hal, pemberian balikan inilah yang sebenarnya dapat mendorong siswa untuk belajar lebih baik, dan oleh karenanya kepada guru perlu ditekankan semangat ini untuk meningkatkan kualitas pembelajaran.

Namun demikian, beban-beban yang terasa memberatkan tersebut akan terbayar jika ternyata penguasaan matematika siswa dapat meningkat dengan baik.

Di sisi lain, berdasar angket yang diberikan kepada siswa dan wawancara guru dengan siswa diperoleh kesan bahwa hampir seluruh siswa merasa mendapat balikan ketika mengerjakan soal tahap I dan soal tahap II. Hampir seluruh siswa juga merasa cukup terbantu atau sangat terbantu ketika mendapatkan balikan dari guru. Lebih dari 80% siswa merasa kesenangannya terhadap matematika bertambah dengan diterapkannya AfL.

Tabel 9.5. Rerata Nilai Siswa pada Soal tahap I, Soal tahap II, Sub-sumatif Sebelum dan Setelah Pelaksanaan AfL, dan Kelas Lain yang Tidak Memakai AfL

No	Nama SMP	Nilai Pengerjaan Soal Tahap I			Nilai Pengerjaan Soal Tahap II			Nilai Sub-suma-tif sebe-lum AfL	Nilai Sub-suma-tif sete-lah AfL	Nilai Subsu-matif kelas pemban ban-ding
		Pembelajaran ke-			Pembelajaran ke-					
		1	2	3	1	2	3			
KELAS VII										
1	SMPN 14	88.6	63.1	74,3	69,1	85,4	94,3	67,3	84,0	61,86
2	SMPN 18	57.5	58.6	58,1	61,8	73.1	66.0	39,5	51,4	48,54
3	SMPN 19	59.7	63.9	79.8	70.5	64,9	92,1	70,1	73,4	65.77
4	SMPN 20	56.6	75.0	73,2	69,5	81,7	74,5	61,8	67.6	56,60
5	SMP Mh1	80,2	86,7	89,2	85,3	81,9	87,5	61,5	67,3	63,38
6	SMP KK	66,2	76,9	84,8	66,2	81,5	86,2	59,4	75,9	72,50
Rerata Besar		68,1	70,7	76,6	70,4	78,1	83,4	59,9	69,9	61,44
KELAS VIII										
1	SMPN 14	84,2	76,3	73,9	75,8	73,4	78,4	66,4	77,7	67,37
2	SMPN 18	75,8	80,7	94,4	60,2	81,9	82,3	50,6	59,4	60,94
3	SMPN 19	87,8	83,3	91,9	85,5	90,2	96,3	67,2	73,1	67,13
4	SMPN 20	87,5	73,3	77,5	86,1	77,6	91,9	66,8	75,9	63,95
5	SMP Mh1	64,6	77,6	89,5	84,3	78,8	89,3	60,8	84,2	61,60
6	SMP KK	82,8	71,2	86,2	81,7	95,3	86,0	74,0	84,8	74,25
Rerata Besar		80,5	77,1	85,8	78,9	82,9	87,2	64,3	75,9	64,21
Kelas IX										
1	SMPN 14	80.9	86.3	87,3	83,7	93,5	97.8	59.9	66.7	49,91
2	SMPN 18	65,8	92,6	57,3	73,0	83,3	93,2	41,0	54,3	56,41
3	SMPN 19	86.8	86,4	97,0	96,6	90,1	81,2	73,8	83,6	71,60
4	SMPN 20	95,8	90,4	87,1	99,0	86,4	87,4	65,2	85.6	67.26
5	SMP Mh1	76,7	87,3	78,7	87,7	92,0	84,0	56,9	72,4	69,67
6	SMP KK	70,6	87.6	85,0	85,0	96,5	98,1	70,4	87,5	71,97
Rerata Besar		79.4	88.4	82,1	87,5	90,3	90,3	61,2	75,1	64,47

Pembandingan Nilai Sebelum AfL dan Sesudah AfL

Tabel 9.5 memuat rerata nilai-nilai siswa pada soal tahap I. nilai siswa pada soal tahap II. nilai sub-sumatif sebelum dan setelah pelaksanaan AfL. dan nilai siswa kelas lain yang tidak memakai AfL di sekolah tertentu.

Berdasarkan Tabel 9.5 dapat dilihat bahwa ada peningkatan yang cukup tajam nilai-nilai siswa sebelum dan sesudah pelaksanaan AfL, dari 59,9 ke 69,9 pada kelas VII, dari 64,3 ke 75,9 pada kelas VIII, dan dari 61,2 ke 75,1 pada kelas IX.

Di sisi lain, walaupun tidak secara menyeluruh, terdapat kecenderungan bahwa nilai siswa pada pembelajaran ketiga lebih baik daripada nilai siswa pada pembelajaran kedua dan nilai siswa pada pembelajaran kedua lebih baik daripada nilai siswa pada pembelajaran pertama.

Pembandingan dengan Kelas Lain yang Tidak Menggunakan AfL

Berdasarkan Tabel 9.5, kecuali untuk kelas VIII dan kelas IX SMP Negeri 18 Surakarta, dapat dilihat bahwa nilai sub-sumatif kelas AfL selalu lebih baik daripada nilai sub-sumatif kelas yang tidak menggunakan AfL. Hal ini menunjukkan bahwa pembelajaran yang mengakomodasi AfL lebih efektif dibandingkan dengan pembelajaran yang tidak mengakomodasi AfL.

Temuan penelitian pada penelitian tersebut di atas menunjukkan bahwa model AfL yang dibangun dapat dilaksanakan dengan baik dan dapat meningkatkan kemampuan matematika siswa. Hal ini sejalan dengan temuan penelitian pada penelitian Mansyur (2009), Young (2005), dan Stiggins & Chappuis (2006).

PENILAIAN OTENTIK (*AUTHENTIC ASSESSMENT*)

Seperti pada penilaian berbasis kelas, ada beberapa definisi mengenai penilaian otentik yang dikemukakan para ahli, yang tidak seluruhnya koheren. Oleh karena itu, pembaca diminta berhati-hati untuk memaknai apa yang disebut penilaian otentik.

Penilaian otentik adalah jawaban terhadap kritik bahwa penilaian yang dilakukan pendidik kebanyakan adalah *paper and pencil test* yang lebih berorientasi kepada pengujian pengetahuan siswa yang bersifat kognitif dan/atau teoretis. Untuk memulai diskusi, perhatikan butir soal pada Contoh 9.1 Contoh 9.2, Contoh 9.3, dan Contoh 9.4 berikut.

Contoh 9.1

Tulislah cara-cara orang menanam jagung!

Contoh 9.2

Sediakan alat-alat yang diperlukan untuk menanam jagung. Kemudian. tanamlah jagung di lahan yang sudah disediakan.

Contoh 9.3

Carilah akar-akar persamaan $5 + \frac{6}{x} + \frac{1}{x^2} = 0$!

Contoh 9.4

Rokok A yang harga belinya Rp10.000.00 dijual dengan harga Rp11.00.00 per bungkus, sedangkan rokok B yang harga belinya Rp15.00.00 dijual dengan harga Rp17.00.00 per bungkus. Seorang pedagang rokok yang mempunyai modal Rp3.000.000.00 dan kiosnya dapat menampung paling banyak 250 bungkus rokok. Berapa laba maksimum yang diperoleh pedagang itu dengan hanya menjual dua jenis rokok tersebut?

Butir soal pada Contoh 9.1 merupakan butir soal yang menanyakan pengetahuan peserta tes dalam menanam jagung. Walaupun seseorang dapat menjawab dengan baik butir soal tersebut, namun belum tentu sorang tersebut dapat menanam jagung dengan baik. Butir soal seperti ini lah yang dikritik sebagai butir soal yang tidak otentik. Sering disebut sebagai penilaian tradisional (*traditional assessment*).

Bandingkan dengan suruhan pada butir soal pada Contoh 9.2. Butir soal seperti pada Contoh 9.2 itulah yang disebut dengan penilaian otentik. Pengertian penilaian otentik seperti itu cocok dengan definisi penilaian otèntik dari Callison (1988) yang mengatakan bahwa "*authentic assessment ia an evaluation process that involves multiple forms of performance measurement reflecting the student's learning, achievement, motivation, and attitudes on instructionally-relevant activities*". Pada definisinya Callison, penilaian otentik adalah penilaian kinerja (*performance*) yang mengandalkan gerak psikomotor. Lebih tegas lagi, Mueller (2005) mengatakan bahwa penilaian otektik adalah "*a form of assessment in which students are asked to perform real-world tasks that demonstrate meaningful application of essential knowledge and skills*".

Berdasarkan dua pengertian tersebut maka dapat dikatakan bahwa penilaian otentik adalah penilaian di mana para siswa diminta untuk mendemonstrasikan aplikasi dari pengetahuan dan keterampilan yang diperoleh dalam kelas ke kejadian nyata. Beberapa orang mengatakan penilaian otentik adalah *performance assessment* (penilaian kinerja, sebab lebih menitikberatkan kepada kinerja daripada pengetahuan semata), atau *alternative assessment* (penilaian alternatif, sebab berbeda dengan penilaian tradisional), atau *direct assessment* (penilaian langsung, sebab penilaian otentik menyediakan bukti langsung dari aplikasi pengetahuan).

Dengan melihat definisi Callison dan Mueller di atas, maka butir soal pada Contoh 9.3 dan Contoh 9.4 bukanlah butir soal untuk penilaian otentik. Walaupun untuk menyelesaikan butir soal pada Contoh 9.3 dan Contoh 9.4 diperlukan pengetahuan tinggkat tinggi (analisis, sintesis, dan evaluasi), tetapi butir soal itu tidak terkait dengan tugas nyata di kehidupan sehari-hari dan bukanlah berbasis kepada aspek psikomotor (*performance*).

Penilaian otentik biasanya mencakup serangkaian tugas yang harus dilakukan oleh siswa. Serangkaian tugas tersebut dinilai melalui rubrik yang dengan rubrik tersebut *performance* (unjuk kerja) dari siswa dapat diukur. Menurut Mueller (2005) tugas yang harus dikerjakan tersebut haruslah *real-world tasks*, tugas yang benar-benar banyak dilakukan di dunia nyata, bukan tugas rekaan atau tugas yang seolah-olah.

Penilaian otentik dikembangkan berdasarkan beberapa pemikiran berikut: (1) misi dari sekolah adalah untuk menciptakan warga negara yang produktif (*productive citizens*), (2) untuk menjadi warga negara yang produktif, seseorang harus dapat melakukan pekerjaan yang bermakna di dunia yang real (*performing meaningful tasks in the real world*), (3) oleh karena itu, sekolah harus dapat membantu siswa untuk dapat menyiapkan diri melakukan pekerjaan di dunia nyata setelah mereka lulus, (4) untuk menopang kesuksesan siswa, sekolah harus meminta para siswa untuk melakukan tugas-tugas bermakna yang merupakan replikasi dari *real world challenges* untuk melihat apakah siswa mampu untuk melakukan hal tersebut.

Langkah-langkah untuk melaksanakan penilaian otentik adalah: (1) identifikasikan standard yang harus dipenuhi, (2) pilihlah pekerjaan yang harus dilakukan, (3) identifikasikan kriteria yang harus dipenuhi dalam melakukan pekerjaan itu, dan (4) buatlah rubriknya yang sesuai.

Berdasarkan definisi penilaian otentik dari Mueler tersebut di atas, dapat dipahami bahwa penilaian otentik sangat cocok untuk mata-mata pelajaran yang bersifat vokasi (keterampilan fisik). Penilaian otentik tidak cocok untuk mata pelajaran yang bersifat kognitif, seperti misalnya Matematika. Atau paling tidak sangat sulit untuk membuat butir soal penilaian otentik di Matematika. Contoh penggunaan penilaian otentik di bidang Matematika adalah penilaian yang dilakukan oleh Chance (1997) pada mata kuliah Pengantar Statistik. Tugas-tugas yang diberikan kepada mahasiswa sebagai wujud dari penilaian otentik yang dilakukannya adalah meminta mahasiswa untuk mengumpulkan data real dari lapangan, men-gorganisasikan, mengolah, dan melaporkan hasil analisisnya.

Definisi lain mengenai penilaian otentik diberikan oleh Winograd & Perkins (1996) sebagai berikut.

Authentic assessment is assessment that occurs continually in the context of a meaningful learning environment and reflects actual and worthwhile learning experiences that can be documented through observation, anecdotal records, journals, logs, work samples, conferences, portfolios, writting discussions, experiments, presentations, exhibits, project and other methods".

Berbeda dengan definisi Callison dan Mueller, definisi Winograd dan Perkins bersifat agak umum. Melalui cara dokumentasinya dapat dilihat bahwa penilaian otentik tetap bukan penilaian yang bersifat kognitif semata, tetapi penilaian yang tetap menampilkan keterampilan psikomotor. Butir soal seperti pada Contoh 9.3 dan Contoh 9.4 tetap bukan merupakan penilaian otentik.

Senada dengan definisi Winograd dan Perkins, O'Maley (Callison, 1998) bahwa karakteristik *sudent performance* yang dapat dianggap sebagai ciri-ciri penilaian otentik adalah sebagai berikut.

- ***Constructed response**: The students contructs responses based on experiences he or she brings to the situation and new multiple reseources are explored in order to create a product.* (siswa memberikan argumentasi berdasarkan pengalaman untuk memperoleh situasi dan sumber baru yang dieksplorasi agar dapat menciptakan suatu produk).
- ***Higher-Order Thinhking**: Responses are made to open ended questions that require skills in analysis, synthesis, and evaluation.* (jawaban dibuat berdasakan pertanyaan terbuka dalam aspek analisis, sintesis, dan evaluasi).
- ***Authenticity**: Tasks are meaningful, chalenging, and engaging activities that mirror good instruction often relevant to a real world context.* (tugas yang diberikan berkmakna, menantang, dan mendukung kegiatan-kegiatan yang mencerminkan pembelajaran yang baik yang relevan dengan konteks kehidupan nyata).
- ***Integrative**: Tasks call for a combination of skills that integrate language arts with other content across the curicullum with all skills and content open to assessment.* (tugas memerlukan kombinasi dari berbagai keahlian dan hal-hal yang terbuka).
- ***Process and Product**: Procedures and strategies for deriving potential responses and exploring multiple solutions to complex problems are often assessed in adition to or in place of a final product or single-correct-response.* (prosedur dan strategi yang menimbulkan respons potensial dinilai bersama-sama dengan produk final atau jawaban benar yang tunggal).
- ***Depth in Place of Breads**: Performance assessment build over time with varied activities to reflect growth, maturity, and depth, leading to mastery of strategis and processes for solving problems in spesific areas with the assumption that these skills will transfer to solving other problems.*

(penilaian otentik/kinerja dibangun bersama-sama dengan kegiatan-kegiatan yang mencerminkan pertumbuhan, kedewasaan, dan kedalaman yang menuju kapada strategi dan proses untuk menyelesaikan masalah dengan asumsi bahwa keterampilan-keterampilan tersebut dapat dikenakan pada permasalahan yang lain).

Menurut O'Malley dan Peirece (Callison, 1998), contoh-contoh penilaian otentik adalah sebagai berikut.

- ***Oral Interviews:*** *Teacher asks student questions about personal background, activities, readings, and other interests.* (guru menanyakan latar belakang pribadi, kegiatan-kegiatannya, bacaan, dan hal-hal lain yang disenanginya).
- ***Story or Text Retelling:*** *Students retells main ideas or selected details of text experienced through listening or reading.* (siswa menceritakan kembali apa ide pokok atau ide terpilih berdasarkan kegiatan mendengarkan atau kegiatan bercerita).
- ***Writing Samples:*** *Student generate narrative , expository, persuasive, or reference paper.* (siswa menulis makalah yang naratif, menjelaskan dan persuasif).
- ***Projects/Exhibitions:*** *Student works with other students as a team to create a project that often involves multimedia production, oral, and written presentations, and a display.* (siswa bekerja dangan siswa lainnya untuk membuat proyek yang sering melibatkan produksi multimedia, presentasi lisan atau tulis, dan pameran).
- ***Experiments/Demonstrations:*** *Student documents a series of experiments, illustrated a procedure, performs the necessary steps to complete a tasks, and documents the results of the actions.* (siswa mendokumentasikan serangkaian eksperimen, menjelaskan prosedur, melakukan langkah-langkah yang diperlukan dan mendokumentasikan hasil kegiatannya).
- ***Constructed-Response Items:*** *Student responds in writing to open ended questions.* (siswa menjawab secara tertulis soal-soal terbuka yang diberikan).

Seiring dengan berkembangnya waktu, definisi penilaian otentik kadang-kadang menyimpang dari apa yang dikatakan oleh Mueller di atas. Banyak pakar mendefinisikan penilaian otentik menurut pemikirannnya sendiri, sampai-sampai Whitelock & Cross (2012) mengatakan bahwa "*authentic assessment is not only a difficult notion to define but it is also problematic to collate features within an assessment task that define it as authentic assessment.*" Ia mengakui bahwa sulit untuk mendefinisikan penilaian otentik yang disepakati oleh semua orang dan sulit untuk mengatakan ciri-ciri penilaian otentik. Namun demikian, setelah ia mempelajari berbagai definisi penilaian otentik dari banyak pakar, ia menyimpulkan bahwa kebanyakan pakar mencirikan penilaian otentik sebagai berikut:

- *collaboration, that is that experienced by practitioner or experts in the field:* (adanya kerjasama, seperti yang dilakukan oleh para praktisi atau ahli di lapangan).
- *simulation of role-play or scenarios:* (adanya simulasi dari suatu permainan atau skenario).
- *problem tasks that are like those encountered by practioners or experts in the field;* (adanya tugas yang seperti yang dialami oleh praktisi atau ahli di lapangan).
- *resources (documents, data, etc) taken spesifically from real world case studies or research;* (sumber-sumber yang ditelaah, misalnya dokumen atau data, diambil dari lapangan (dunia nyata) atau dari suatu riset).
- *tasks that students find meaningfull;* (siswa merasa bahwa tugas itu berarti atau bermakna).
- *examinations taking place in real word settings;* (ujian-ujian dilaksanakan seperti yang terjadi di dunia nyata).
- *a range of assessment tasks rather than just 'traditional' ones;* (tugas sebagai wujud dari penilaiannya tidak lagi tradisional).
- *demonstration and use of judgement;* (menunjukkan dan menggunakan justifikasi).
- *students being involved in the negotiation of the assessment task;* (siswa dilibatkan dalam penentuan tugas).
- *a test of how well the student thinks like practitioner/expert in the field (i.e.* 'intune' with the 'disclipinary mind'); (tes untuk melihat apa yang siswa pikirkan seperti cara pemikiran praktisi atau ahli di lapangan).

Kurikulum 2013 melalui Permendikbud Nomor 66 Tahun 2013 tentang Standar Penilaian Pendidikan menyatakan bahwa penilaian otentik adalah penilaian yang dilakukan secara komprehensif untuk menilai mulai dari masukan (*input*), proses, dan keluaran (*output*) pembelajaran. Di sisi lain, Permendikbud Nomor 104 Tahun 2014 tentang Penilaian Hasil Belajar oleh Pendidik menyatakan bahwa penilaian otentik adalah "penilaian yang menghendaki peserta didik menampilkan sikap, menggunakan pengetahuan, dan keterampilan yang diperolehnya dari pembelajaran dalam melakuan tugas pada situasi yang sesungguhnya". Permendikbud Nomor 104 Tahun 2014 juga mengatakan bahwa "soal tes tertulis yang menjadi penilaian otentik adalah soal yang menghendaki peserta didik merumuskan jawabannya sendiri, seperti soal-soal uraian". Menggunakan definisi penilaian otentik pada Kurikulum 2013, bisa jadi butir soal Contoh 9.3 dan Contoh 9.4 merupakan penilaian otentik[1], yang menurut Mueller (1005) dan Winograd & Perkins (1996) bukanlah penilaian otentik. Menurut Kurikulum 2013, tampaknya, yang bukan merupakan penilaian otentik adalah penilaian yang dinyatakan dalam bentuk pilihan ganda seperti misalnya pada Ujian Nasional.

---

[1] Menurut penulis, definisi penilaian otentik di Kurikulum 2013 tidak benar-benar jelas.

Sebagai catatan mengenai penilaian otentik ini, sebaiknya jangan mengharapkan penilaian otentik dapat dengan mudah diterapkan pada semua mata pelajaran. Penilaian otentik sangat sukar diterapkan pada pelajaran yang lebih bersifat kognitif. seperti Matematika, melainkan mudah diterapkan pada mata pelajaran di SMK bidang keahlian Tata Busana dan Tata Boga.

BAHAN DISKUSI

1. Dikenal adanya dua penilaian, yaitu penilaian formatif dan penilaian sumatif. Manakah yang lebih membantu siswa belajar, penilaian formatif atau penilaian sumatif?

2. Perhatikan penilaian yang dilakukan oleh guru atau dosen Anda. Manakah yang lebih banyak dilakukan oleh guru dan dosen Anda, penilaian formatif atau sumatif?

3. Di Kurikulum 2013 dikenal adanya ulangan harian, yang didefinisikan sebagai penilaian yang dilakukan setiap menyelesaikan satu muatan pembelajaran.
   a. Dapatkah ulangan harian berfungsi sebagai penilaian formatif? Pada keadaan seperti apa?
   b. Dapatkah ulangan harian berfungsi sebagai penilaian sumatif? Pada keadaan seperti apa?

4. Seorang guru memberikan ulangan harian setelah selesai membelajarkan satu muatan tertentu. Guru itu lalu memeriksa pekerjaan siswa-siswanya, memberi nilai, dan menyimpan nilai itu untuk menentukan nilai rapor akhir semester. Tidak ada komentar apapun mengenai pekerjaan siswa. Guru hanya memberi nilai saja.
   a. Apakah diperkenankan seorang guru memberi ulangan harian seperti itu? Mengapa?
   b. Pelaksanaan ulangan harian seperti merupakan penilaian formatif atau sumatif? Mengapa?

5. Di tengah-tengah pembelajaran, seorang guru Matematika memberikan soal pendek untuk dikerjakan di kelas selama 10 menit. Setelah selesai pengerjaan, setiap murid diminta untuk menilai dan memberi komentar terhadap pekerjaan temannya dengan menggunakan rubrik penilaian rinci yang dibuat oleh gurunya. Penilaian semacam itu adalah salah satu contoh *peer-assessment* (penilaian teman sejawat).
   a. Apakah penilaian teman sejawat tersebut dapat dikatakan sebagai *assessment for learning* (AfL)? Mengapa?

b. Apakah penilaian teman sejawat tersebut dapat dikatakan sebagai penilaian formatif? Mengapa?
c. Apakah menurut Anda, penilaian teman sejawat tersebut perlu dilakukan oleh guru di kelas? Mengapa?

6. Di akhir suatu pembelajaran, seorang guru Matematika memberikan soal pendek untuk dikerjakan di kelas selama 10 menit. Setelah selesai pengerjaan, setiap murid diminta untuk menilai dan memberi komentar terhadap pekerjaan temannya dengan menggunakan rubrik penilaian rinci yang dibuat oleh gurunya. Penilaian semacam itu juga merupakan contoh *peer-assessment* (penilaian teman sejawat).
   a. Apakah penilaian semacam itu merupakan AfL? Mengapa?
   b. Apakah penilaian semacam itu dapat disebut sebagai penilaian sumatif? Mengapa?

7. Pada Kurikulum 2013 dikenal adanya ulangan harian, ulangan tengah semester, dan ulangan semester. Ulangan harian adalah penilaian yang dilakukan setiap menyelesaikan satu muatan pembelajaran. Ulangan tengah semester adalah penilaian yang dilakukan untuk semua muatan pembelajaran yang diselesaikan dalam paruh pertama semester Ulangan akhir semester adalah penilaian yang dilakukan untuk semua muatan pembelajaran yang diselesaikan dalam satu semester.
   a. Dari ketiga penilaian itu, manakah yang merupakan penilaian berbasis kelas? Mengapa?
   b. Dari ketiga penilaian itu, manakah yang merupakan AfL? Mengapa?
   c. Dari ketiga penilaian itu, manakah yang merupakan penilaian otentik? Mengapa?

8. Apakah setiap soal bentuk uraian merupakan wujud dari penilaian otentik? Mengapa?

9. Misalnya seorang siswa taman kanak-kanak diminta untuk menceritakan cita-citanya jika ia dewasa. Apakah seperti itu merupakan penilaian otentik?

10. Setujukah Anda terhadap pendapat bahwa penilaian otentik adalah penilaian yang tidak dalam bentuk benar-salah, atau menjodohkan, atau pilihan ganda? Mengapa?

***

BAB X
PENILAIAN PORTOFOLIO

PENDAHULUAN

Penilaian portofolio relatif baru dalam pengukuran pendidikan. Namun demikian, penilaian ini banyak menarik perhatian para pendidik, karena penilaian ini memberikan alternatif lain dalam penilaian pembelajaran.

Portofolio adalah kumpulan kerja (prestasi) seseorang yang tersusun secara sistematis. Dalam pembelajaran, portofolio merujuk kepada kumpulan sistematis kerja atau karya siswa. Pada kenyataannya, portofolio merupakan metode yang bagus bagi para profesional untuk menunjukkan keterampilan dan kemampuannya. Dalam bidang fotografi, misalnya, kumpulan foto-foto seorang fotograf yang dipamerkan akan menunjukkan seberapa profesional fotografer tersebut. Dalam bidang seni lukis, misalnya, kumpulan lukisan yang dipamerkan oleh seseorang akan menunjukkan seberapa tinggi kemampuan pelukis yang bersangkutan. Pada kasus seperti ini, portofolio adalah metode yang paling bagus untuk menunjukkan keterampilan dan keahlian seseorang. Fitur penting portofolio adalah bahwa portofolio harus terbarui seiring dengan pertumbuhan keterampilan dan kemampuan seseorang.

PEMAKAIAN DI KELAS

Para pendukung penilaian portofolio percaya bahwa hubungan antara pembelajaran dan penilaian dapat diperkuat sebagai konsekuensi dari akumulasi kerja siswa yang terus menerus dalam portofolionya. Secara ideal, guru yang menggunakan portofolio dalam pembelajarannya akan meletakkan *ongoing collection and appraisal students' work* sebagai sentral dari program pembelajarannya dibandingkan dengan *peripheral activity* di mana hanya secara sekali-kali guru mengumpulkan data untuk meyakinkan

pengawas atau orang tua murid bahwa segala sesuatu telah berjalan baik di dalam kelas.

Berikut ini diberikan contoh penggunaan portofolio untuk menilai kemajuan siswa dalam mata pelajaran Bahasa Indonesia. dalam hal ini menulis karangan. Si Guru, misalnya namanya Pak Pholio, meminta para siswanya untuk menyimpan tiga portofolio. Pada setiap portofolio, para siswa diminta menyimpan pekerjaan mereka dan perbaikannya setelah mendapatkan balikan. Setiap pekerjaan diberi tanggal, sehingga Pak Pholio dan siswa dapat melihat seberapa jauh perbedaan kualitas terjadi pada sepanjang waktu. Asumsinya, jika pembelajaran efektif dapat berlangsung, pasti dapat dilihat peningkatan kualitas siswa dalam menulis karangan.

Tiga atau empat kali dalam satu semester, Pak Pholio mengadakan presentasi portofolio selama 15 – 20 menit untuk setiap orang mengenai masing-masing portofolionya. Selama presentasi, guru dan siswa yang bersangkutan melakukan penilaian terhadap hasil kerjanya. Menjelang berakhirnya semester, siswa-siswa diminta memamerkan hasil kerjanya, tidak saja hasil kerjanya yang terbaik, tetapi juga cara mendapatkannya. Pameran ini dipajang di tempat tertentu, sehingga orang tua siswa yang berkunjung ke sekolah dapat melihatnya. Pak Guru juga dapat mengirimkan portofolio siswa, jika orang tua tidak dapat berkunjung ke sekolah.

Salah seorang tokoh pembelajaran dan penilaian di bidang seni, Roger Farr, mengatakan bahwa *the real payoff from proper portofolio assessment is that students' self-evaluation capabilities are enhanced*. Jadi, selama *konferensi* portofolio, guru men-*encourage* siswa untuk dapat menilai karyanya sendiri. Kecuali itu, kemampuan siswa untuk menilai dirinya sendiri dikembangkan, tidak saja pada *konferensi* portofolio, tetapi juga selama berlangsungnya pembelajaran di sekolah.

Untuk tujuan evaluasi diri, para siswa diminta untuk membandingkan hasil kerja semula dengan hasil kerja berikutnya. Evaluasi diri ini dianggap sangat berguna, baik dari sisi perpektif pembelajaran, tetapi juga pada masa depan kehidupan mereka.

Penilaian portofolio, di samping dilakukan oleh siswa sendiri, tentu saja juga dilakukan oleh guru.

Penilaian portofolio sangat bagus dikenakan untuk mata-mata pelajaran tertentu, misalnya Bahasa Indonesia. Pada mata pelajaran ini, para peserta didik dapat diminta mengumpulkan portofolionya yang berupa hasil kerjanya (misalnya puisi dan cerpen) di majalah-majalah tertentu. Penilaian portofolio tentu saja sangat bagus untuk mata-mata pelajaran keterampilan di SMK, misalnya Tata Busana. Pada waktu-waktu tertentu siswa dapat memamerkan hasil karyanya di hadapan para siswa lainnya.

LANGKAH-LANGKAH PENILAIAN PORTOFOLIO DI KELAS

Berikut ini diberikan beberapa langkah penting untuk melakukan penilaian portofolio.

1. Agar portofolio dapat menyatakan perkembangan kerja siswa secara akurat dan untuk memperkuat semacam penilaian diri yang sangat krusial dalam portofolio, siswa harus diberi penjelasan bahwa portofolio adalah kumpulan dari kerja mereka sendiri dan bukan semata-mata wadah kerja siswa yang akan dinilai oleh gurunya. Dalam konteks ini, guru dapat menjelaskan fungsi yang berbeda dari biasanya dalam penilaian portofolio.
2. Berbagai jenis kerja siswa dapat dipilih untuk dimasukkan ke dalam suatu portofolio. Yang lebih bagus adalah guru dan siswa berunding untuk menentukan kerja apa yang akan dimasukkan ke dalam portofolio.
3. Siswa diminta untuk mengumpulkan hasil kerjanya dalam suatu wadah yang baik dan menempatkannya ke dalam suatu tempat yang baik, misalnya *file cabinet*. Guru dapat membantu siswa memilih kerja siswa yang harus dimasukkan ke dalam portofolio.
4. Guru bersama-sama dengan siswa menentukan kriteria untuk menentukan kualitas portofolio. Barangkali hal ini bukanlah sesuatu yang mudah dikerjakan, karena portofolio lebih bersifat individual.
5. Dengan menggunakan kriteria yang telah disetujui, siswa dapat diarahkan untuk mengevaluasi kerjanya, baik secara holistik maupun secara analitik, atau kombinasi di antara keduanya. Semacam penilaian diri dapat dibuat secara rutin dengan menggunakan kartu, misalnya, yang mengidentifikasi kekuatan dan kelemahan kerja siswa serta melakukan usulan bagaimana kualitas kerja itu dapat ditingkatkan.
6. Tukar pikiran antara guru dan siswa mengenai hasil kerja siswa merupakan aspek penting untuk meyakinkan bahwa penilaian portofolio memenuhi tujuannya. Presentasi tidak saja berfungsi untuk menilai kerja siswa, tetapi juga untuk meningkatkan kemampuan penilaian diri siswa. Adakan presentasi portofolio sebanyak mungkin. Agar supaya presentasi portofolio efisien, usahakan agar siswa benar-benar siap untuk menyajikan *the topics of most concern* baik untuk guru maupun siswa.
7. Guru harus memberi tahu kepada orang tua siswa untuk mengerti the *nature of the portofolio assessment process*. Diharapkan orang tua siswa dapat pula mengikuti perkembangan siswa. Semakin aktif orang tua ikut mereview hasil kerja siswa, semakin kuat pesan penilaian portofolio untuk kemajuan anaknya.

KEUNGGULAN-KEUNGGULAN PENILAIAN PORTOFOLIO

Menurut Reynolds, Livingstone, dan Willson (2010: 271), penilaian portofolio mempunyai sejumlah keunggulan sebagai berikut.

Portofolio sangat bagus untuk menunjukkan prestasi siswa dan perkembangannya dari waktu ke waktu. Dalam keahlian melukis, misalnya kumpulan lukisan seseorang akan menunjukkan perkembangan kematangan melukisnya dari waktu ke waktu. Dalam keahlian menulis cerita pendek dan/atau artikel di koran, seseorang juga akan dapat melihat perkembangan kematangan menulis.

Portofolio dapat meningkatkan kemampuan siswa dalam suatu hal dan pada akhirnya meningkatkan prestasi dan produk mereka. Karya-karya yang dimasukkan ke dalam portofolio dipilih oleh siswa sendiri dan akan dipamerkan kepada orang lain, sehingga hal ini dapat menimbulkan motivasi dan kehendak untuk terus belajar lebih giat.

Untuk siswa yang produktif, portofolio dapat dipakai sebagai sarana untuk melakukan penilaian terhadap karya-karyanya. Dalam jangka panjang hal ini akan memupuk kemampuan siswa untuk dapat melaksanakan dan meningkatkan *self-assessment skills*.

Jika digunakan secara baik, portofolio dapat meningkatkan keterkaitan antara pembelajaran dan penilaian. Pada waktu-waktu tertentu, portofolio siswa dapat dipamerkan dan didiskusikan di kelas. Dalam kasus seperti ini, maka pembelajarannya menyatu dengan penilaian portofolio itu sendiri.

Guru yang baik harus selalu mengikuti kemajuan belajar siswanya. Dalam kasus portofolio, guru yang baik harus selalu mengikuti perkembangan hasil karya siswanya dan memberikan komentar atas karya-karya tersebut. Oleh karena itu, melalui portofolio akan terjadi komunikasi yang baik antara guru dan siswa.

KELEMAHAN-KELEMAHAN PENILAIAN PORTOFOLIO

Pada penilaian portofolio, seperti halnya pada *constructed-response measurement*, sulit dilakukan penilaian secara masal, sebab penilaian portofolio pada dasarnya harus merupakan keinginan, minat, dan kemampuan individual siswa. Terkait dengan ini, pada penilaian portofolio juga sulit untuk membuat kriteria penilaian yang dapat mengakomodasi semua kerja siswa.

Pelaksanaan penilaian portofolio menyita banyak waktu, sehingga untuk guru yang sangat sibuk akan sulit untuk dapat melakukan penilaian portofolio dengan baik. Namun demikian, pendukung penilaian portofolio meyakinkan bahwa waktu yang terpakai untuk melakukan penilaian

portofolio akan terbayar dengan tumbuhnya evaluasi diri siswa yang kelak akan berguna bagi hidupnya.

Pada dasarnya penilaian portofolio terkait dengan karya-karya (melukis, menulis, memahat, membuat produk, dan semacamnya), yang oleh karenanya terkait erat dengan aspek psikomotor. Dengan demikian, penilaian portofolio tidak dapat dengan baik diterapkan pada mata pelajaran yang bersifat kognitif, seperti Matematika. Namun demikian, sering kali diharapkan portofolio dapat diterapkan untuk semua mata pelajaran. Sehingga oleh karenanya, wujud portofolio di mata pelajaran Matematika adalah kumpulan pekerjaan siswa dalam mengerjakan soal, yang hal ini sebenarnya menyimpang dari pengertian portofolio semula.

Kelemahan menonjol pada penilaian portofolio adalah sukar dalam penskorannya. Pada kasus ini sukar untuk menentukan rubrik penskorannya yang adil. Kecuali itu, unsur subjektivitas penilai sangat menonjol dalam penilaian portofolio.

RUBRIK PENILAIAN PORTOFOLIO

Berikut ini adalah contoh rubrik penilaian portofolio pada mata pelajaran Bahasa Indonesia yang diambil dari Permendikbud Nomor 104 Tahun 2014 , sebagai berikut.

Contoh 10.1

Nama Siswa:

No	Kompetensi Dasar	Periode	Aspek yang DInilai				Keterangan
			Tata Bahasa	Kosa Kata	Gagasan	Sistematika	
1	Menulis karangan deskriptif						
2	Membuat resensi buku						

BAHAN DISKUSI

1. Menurut Anda, dapatkah penilaian portofolio dipakai pada Ujian Nasional? Mengapa?

2. Menurut pendapat Anda, apakah penilaian portofolio dapat diterapkan untuk semua mata pelajaran? Mengapa?

3. Rencanakan penilaian portofolio, jika memungkinkan, untuk matapelajaran:
   a. Matematika
   b. Sejarah
   c. Ilmu Pengatahuan Alam
   d. Ilmu Pengetahuan Sosial
   e. Bahasa Indonesia
   f. Bahasa Inggris

4. Pada awal tahun dua-ribuan, para guru diminta untuk mengumpulkan portofolio dalam rangka sertifikasi pendidik. Apa yang dikumpulkan oleh guru tersebut? Apakah kegiatan itu juga merupakan kegiatan penilaian portofolio? Jelaskan!

5. Menurut pendapat Anda, apakah guru wajib mempelajari dan menerapkan penilaian portofolio di kelasnya? Mengapa?

****

BAB XI
DIFFEENTIAL ITEM FUNCTIONING

PENDAHULUAN

Riset mengenai bias butir (*item bias*) telah dimulai sejak tahun 1910 oleh Alfred Binet ketika dia berpikir bahwa beberapa butir soal inteligensi yang dibuatnya mungkin lebih mengukur efek pengalaman kultural daripada mengukur kapasitas mental. Riset yang sama dilakukan oleh Wiliam Stern pada tahun 1912 yang menyelidiki perbedaan kelompok pada tes inteligensi di Jerman. Stern menyelidiki sebab-sebab yang mungkin muncul pada perbedaan hasil suatu tes inteligensi (Camilli & Shepard, 1994: 4).

Walaupun telah dimulai sejak awal abad ke-20, oleh Binet dan Stern, studi mengenai bias butir pada suatu tes pertama kali dilakukan secara sungguh-sungguh baru pada tahun seribu sembilan ratus enampuluhan (Angoff, 1993: 3). Studi tersebut didesain untuk mengembangkan metode yang mempelajari adanya perbedaan budaya dan menyelidiki pernyataan yang mengemukakan bahwa sebab utama perbedaan antara siswa kulit hitam dan kulit putih di Amerika Serikat pada tes kemampuan kognitif adalah bahwa tes memuat butir-butir soal yang berada di luar wilayah budaya minoritas. Asumsi awalnya adalah bahwa butir-butir soal berkaitan dengan materi yang siswa-siswa kelompok minoritas mempunyai kesempatan yang lebih kecil untuk mempelajarinya. Tujuan spesifik dari studi ini adalah untuk mengidentifikasikan butir-butir soal yang bias terhadap siswa-siswa kelompok minoritas dan kemudian membuangnya dari tes.

Bias tidak dihasilkan dari kekeliruan random pengukuran. Tidak ada tes yang secara sempurna mengukur *trait* atau domain pengetahuan yang dimaksudkan untuk diukur. tetapi sepanjang kekeliruan pengukuran mengenai anggota-anggota kelompok yang berbeda secara sama, tes tersebut tidaklah bias. Konsep bias butir juga harus dibedakan dengan *adverse impact*. *Adverse impact* atau perbedaan rerata antarkelompok tidak dengan sendirinya merupakan bukti adanya bias. Perbedaan rerata antarkelompok dapat disebabkan karena kemampuan kelompok yang pertama secara keseluruhan lebih baik daripada kemampuan kelompok yang kedua, namun pada anggota-anggota kelompok yang kemampuannya sama dapat saja tidak terdapat perbedaan. Hal yang demikian bukanlah suatu bias dalam suatu tes.

Terdapat dua pendekatan statistik untuk mendeteksi adanya bias tes. Pertama, adalah dengan berdasarkan kriteria yang ada di luar tes dan yang kedua adalah dengan berdasarkan kriteria internal yang ada pada tes. Dua pendekatan tersebut dijelaskan pada paragraf-paragraf berikut.

Berbagai prosedur bias internal dikembangkan dengan menggunakan skor total atau skor sejumlah butir soal di dalam tes sebagai kriteria untuk menentukan perbedaan kelompok. Kemudian, bias diartikan sebagai kesulitan butir relatif yang berbeda pada kelompok yang berbeda. Ide konsep ini adalah memasangkan skor peserta ujian pada skor total untuk melihat apakah pembandingan peserta tes dari kelompok yang berbeda menunjukkan hasil yang sama atau berbeda pada suatu butir soal. Jika berbeda, butir diduga bias.

Sebagai konsekuensi, peneliti yang menyelidiki bias butir dengan kriteria internal memilih menggunakan tes sebagai satu kesatuan sebagai kriteria pengganti untuk memasangkan kemampuan kelompok. Dalam usaha tersebut, disadari sejak semula bahwa studi mengenai bias dengan kriteria internal dapat salah apabila kriteria itu sendiri tidak baik, khususnya jika kriteria itu sendiri bias. Dalam keadaan seperti ini, sangat dimungkinkan butir-butir yang dianggap bias ternyata tidak bias dan sebaliknya butir-butir yang dianggap tidak bias ternyata bias.

Dalam usaha untuk menyelidiki bias butir, berbagai metode dikembangkan oleh para ahli psikometrika untuk menentukan bagaimana butir yang menyimpang (*aberrant*) itu bisa terjadi. Namun

demikian, yang diperoleh dari berbagai studi tersebut hanyalah temuan statistik, yang masih memerlukan adanya interpretasi dan keputusan apakah butir-butir soal yang terdeteksi tersebut merupakan butir-butir soal yang bias atau bukan. Beberapa butir soal yang tidak normal tersebut dapat merupakan butir-butir soal yang memang bias, dalam arti bahwa mereka bertindak tidak adil terhadap kelompok minoritas. Namun demikian, beberapa di antaranya dapat juga diputuskan sebagai butir-butir soal yang adil, dalam arti bahwa hal-hal yang diujikan tersebut merupakan *outcomes* yang penting, cocok untuk semua siswa, tetapi tidak secara sama diketahui dan dimengerti oleh semua siswa.

Dalam keadaan tertentu, pengertian bias cukup jelas dimaknakan, namun kadang-kadang menimbulkan konflik semantik. Kata bias dapat membingungkan ketika pada saat-saat tertentu bias diartikan sebagai perbedaan besar dalam perolehan skor, misalnya kelompok pertama mempunyai rerata skor yang lebih besar daripada rerata skor kelompok kedua. Beberapa usulan dibuat untuk menggunakan istilah lain selain bias dalam kaitannya dengan observasi statistik. Akhirnya, istilah keberbedaan fungsi butir (*differential item functioning, DIF*) digunakan untuk menunjuk kepada hasil observasi bahwa sebuah butir soal berperan berbeda secara statistik pada kelompok yang berbeda. Kemudian, bagaimana butir-butir soal yang terdeteksi tersebut diputuskan dan digunakan, dalam arti apakah terjadi bias secara sosial dan langkah-langkah apa yang akan diambil, adalah persoalan lain yang terpisah.

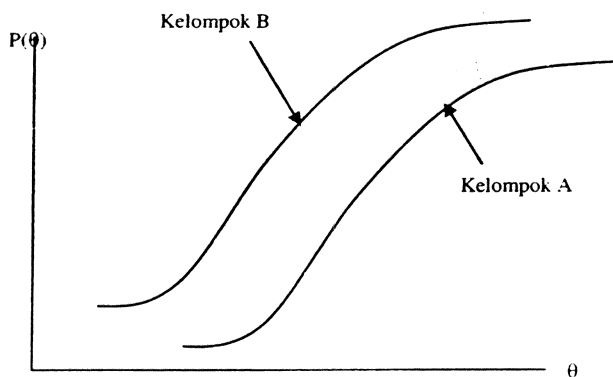
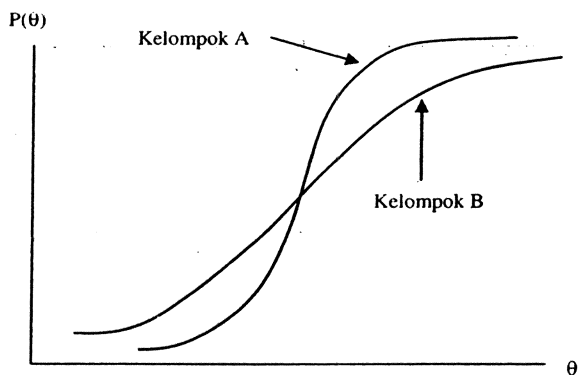
PENGERTIAN *DIF*

Secara konseptual, *DIF* dikatakan muncul pada sebuah butir soal, jika peserta tes yang mempunyai kemampuan yang sama pada konstraks yang diukur oleh tes, tetapi berasal dari kelompok berbeda, mempunyai peluang berbeda dalam menjawab benar butir soal tersebut (Hulin, Drasgow & Parson, 1993: 152, Roussos, Schnipke & Pashley, 1999: 293; Penfield & Lam, 2000: 6). Untuk menentukan apakah suatu butir soal terindikasi *DIF* atau tidak, diperlukan indeks *DIF*, yaitu indeks yang menunjukkan seberapa kuat indikasi *DIF* ada pada butir itu. Jika tingkat indikasi *DIF* tersebut secara praktik dianggap signifikan, dapat dengan mengujinya memakai uji statistik tertentu atau hanya dengan melihat indeksnya saja, maka butir soal

yang bersangkutan dikatakan terkena *DIF*, memuat *DIF*, atau terdeteksi sebagai butir *DIF*.

Dalam konteks teori respons butir, terjadi atau tidak terjadinya *DIF* pada sebuah butir soal terletak kepada fungsi respons butir (*item response function*) untuk butir soal tersebut pada kelompok yang dipersoalkan. Kurva yang menggambarkan fungsi respons butir disebut kurva respons butir atau kurva karakteristik butir (*item characteristic curve, ICC*). Jika sebuah butir soal mempunyai fungsi respons butir yang tepat sama untuk setiap kelompok, maka setiap peserta tes pada setiap kemampuan atau *skill* θ mempunyai peluang yang tepat sama untuk menjawab benar, terlepas dari keanggotaan kelompok. Butir soal yang demikian merupakan butir soal yang tidak memuat *DIF*. Hal ini tetap benar sekalipun suatu kelompok mempunyai rerata θ yang lebih rendah, yang berarti mempunyai skor tes yang lebih rendah dibandingkan dengan skor tes kelompok yang lain. Dalam kasus seperti ini, hasil tes menunjukkan adanya perbedaan kemampuan kelompok dan bukanlah menunjukkan adanya bias. Sebaliknya, jika sebuah butir soal mempunyai fungsi respons butir yang berbeda untuk kelompok yang berbeda, ini merupakan pertanda adanya *DIF* pada butir soal tersebut.

Terdapat dua jenis *DIF*, yaitu *DIF* uniform (konsisten) dan *DIF* tidak uniform (tidak konsisten). *DIF* uniform muncul jika keuntungan salah satu kelompok terhadap kelompok lainnya terjadi pada setiap level kemampuan, sedangkan *DIF* tidak uniform muncul jika keuntungan salah satu kelompok terhadap kelompok lainnya tidak terjadi pada setiap level kemampuan (Penfield & Lam 2000: 9). Jika dikaitkan dengan pengertian interaksi, yang populer pada uji statistik analisis variansi, *DIF* uniform terjadi jika tidak terdapat interaksi antara tingkat kemampuan peserta tes dan keanggotaan kelompok dan *DIF* tidak uniform terjadi jika terdapat interaksi antara tingkat kemampuan peserta tes dan keanggotaan kelompok (Rogers & Swaminathan, 1993: 105).

Terkait dengan teori respons butir, *DIF* uniform terjadi jika kurva karakteristik butir untuk suatu butir soal berbeda untuk kelompok yang berbeda dan kedua kurva tersebut tidak saling berpotongan. Sebaliknya, *DIF* tidak uniform terjadi jika kurva karakteristik butir untuk suatu butir soal berbeda untuk kelompok yang berbeda, namun kedua kurva tersebut berpotongan. Dua situasi *DIF* tersebut ditunjukkan pada Gambar 11.1 dan Gambar 11,2.

Gambar 11.1. Salah Satu Kurva Fungsi Respons Berada di Atas Kurva Fungsi Respons yang Lain

Gambar 11.2. Kedua Kurva Fungsi Respons Berpotongan

Pada Gambar 11.1, kurva karakteristik butir untuk suatu kelompok berada di atas kurva karakteristik butir untuk kelompok lain pada setiap θ. Ini berarti, pada setiap level kemampuan yang sama, peserta tes pada kelompok pertama mempunyai peluang yang lebih baik untuk menjawab benar butir soal tersebut dibandingkan dengan peserta tes pada kelompok kedua. Situasi seperti ini adalah pertanda adanya *DIF* uniform. Butir soal yang kurva karakteristik butirnya seperti pada Gambar 11.1 menguntungkan kelompok B, dan sebaliknya merugikan kelompok A, untuk setiap level kemampuan.

Pada Gambar 11.2, kurva karakteristik butir untuk dua kelompok berpotongan. Ini berarti, pada level kemampuan tertentu, kurva karakteristik butir untuk suatu kelompok berada di atas kurva karakteristik butir untuk kelompok yang lain dan pada level kemampuan yang lain terjadi sebaliknya. Butir soal seperti ini memuat *DIF* yang menguntungkan kelompok pertama pada peserta tes yang mempunyai interval kemampuan tertentu dan menguntungkan kelompok kedua pada peserta tes yang mempunyai interval kemampuan yang lain. *DIF* yang terjadi disebut *DIF* tidak uniform. Pada praktiknya *DIF* tidak uniform jarang terjadi (Camilli & Shepard, 1994: 66).

WAKTU PENDETEKSIAN *DIF*

Seperti disebutkan pada Pendahuluan, pendeteksian *DIF* merupakan bagian esensial dari pengembangan tes. Ini berarti, sebelum tes digunakan, harus dilakukan uji coba untuk mendeteksi keberadaan *DIF*, sama seperti ketika pengembang tes melihat kelayakan nilai paramater butir-butir soal. Dengan cara ini, pengembang tes dapat menganalisis butir-butir soal yang terkena *DIF* dan mengkajinya apakah butir tersebut tetap dipertahankan atau dibuang dari tes.

Namun demikian, seperti yang dikatakan oleh Gierl, Khaliq dan Boughton (1999), ketika melakukan deteksi *DIF* pada ujian Matematika dan Sains di Alberta, tidak semua pengembang tes melakukan deteksi *DIF* pada saat pengembangan tes. Dalam kasus seperti ini, pendeteksian *DIF* setelah tes dilakukan masih tetap berguna untuk pengembangan tes di tahun-tahun berikutnya.

Ketika pendeteksian *DIF* tidak dapat dilakukan pada saat pengembangan tes, Gierl, Khaliq dan Boughton (1999: 16) mengusulkan dua tahapan berikut. Tahap pertama, pengkajian tes (*test review*) dilakukan pada saat pengembangan tes oleh sekelompok pengkaji (*reviewers*), seperti yang telah biasa dilakukan. Tahap kedua, setelah tes dilaksanakan, dengan menggunakan hasil pendeteksian *DIF*, pengkaji yang sama menginterpretasikan butir-butir soal yang terkena *DIF*. Dua tahapan tersebut dapat "*to sensitize developers and item writers to the source of DIF, and to reduce the number of DIF items on a test*" (Gierl, Khaliq dan Boughton, 1999: 16).

TINDAK LANJUT TERHADAP KEBERADAAN *DIF*

Penyebab munculnya *DIF*, yang menunjukkan bahwa suatu butir soal berfungsi berbeda, dapat bermacam-macam, antara lain, karena butir soal tersebut menguntungkan salah satu kelompok karena susunan bahasanya atau karena substansi yang ditanyakan lebih dikenal oleh salah satu kelompok. Penyebab munculnya *DIF* dapat juga karena adanya perbedaan fasilitas antarkelompok, adanya perbedaan kemampuan guru yang mengajar, dan adanya pelaksanaan tes yang tidak adil.

Untuk butir-butir soal yang terkena *DIF* harus dilakukan pembahasan lebih lanjut apakah butir-butir soal tersebut tetap dipakai atau dibuang dari sebuah tes. Jika penyebab munculnya *DIF* tidak terkait dengan konstruks yang diukur oleh tes, misalnya karena menggunakan istilah yang lebih dikenal oleh suatu kelompok dibandingkan dengan kelompok yang lain, maka butir soal yang terkena *DIF* tersebut harus dibuang dari tes. Misalnya sebuah butir soal menanyakan berapa banyaknya roda pada tiga buah becak pada matapelajaran Matematika. Jika butir soal tersebut terkena *DIF* yang menguntungkan peserta tes dari wilayah perkotaan dibandingkan dengan peserta tes dari wilayah pegunungan, maka patut diduga bahwa peserta tes dari wilayah pegunungan tidak dapat menjawab butir soal tersebut karena tidak sering melihat becak. Butir soal yang seperti ini merupakan butir soal yang jelek dan disebut butir soal yang bias. Butir soal yang bias harus dibuang dari tes atau paling tidak direvisi kembali sebelum dipakai.

Di sisi lain, sebuah butir soal yang terkena *DIF* dapat saja masih tetap dipertahankan dalam sebuah tes. Dengan kata lain, butir soal tersebut tetap merupakan butir soal yang tidak jelek, walaupun terkena *DIF*. Namun demikian, tetap diperlukan langkah-langkah lanjutan untuk menghilangkan sumber munculnya *DIF*. Misalnya terdapat sebuah butir soal yang terkena *DIF* yang menguntungkan peserta tes dari wilayah A dan merugikan peserta tes dari wilayah B. Setelah dianalisis substansi yang ditanyakan dan bahasa yang digunakan, ternyata tidak ada unsur yang menguntungkan peserta tes dari wilayah A yang disebabkan oleh faktor substansi dan bahasa. Butir soal yang demikian ini merupakan butir soal yang tidak bias dan tetap harus dipertahankan dalam tes. Penyebab munculnya *DIF* pada butir soal tersebut dapat diduga berasal dari hal-hal di luar tes, misalnya subpokok bahasan yang ditanyakan tidak diajarkan di wilayah B tetapi

diajarkan di wilayah A, proses pembelajaran dan fasilitas pembelajarannya lebih baik di wilayah A daripada di wilayah B, dan semacamnya. Langkah yang harus dilakukan oleh pengambil kebijakan adalah menghilangkan adanya *DIF* pada butir soal tersebut dengan memperbaiki proses pembelajaran di wilayah B dengan harapan pada tahun berikutnya butir soal tersebut tidak terkena *DIF* yang menguntungkan peserta tes dari wilayah A. Dengan sendirinya hal ini dapat dilakukan jika terdapat standardisasi materi yang diajarkan dan terdapat standardisasi kompetensi yang harus dicapai oleh siswa.

Jika karena sesuatu hal, pendeteksian *DIF* baru dapat dilakukan setelah tes dipakai untuk menentukan keputusan dan ternyata, setelah melalui pengkajian, terdapat sejumlah butir soal yang seharusnya dibuang dari tes, maka pengembang tes harus membuang butir-butir tersebut dari bank soal. Dalam kasus seperti ini harus disadari bahwa mungkin terjadi kesesatan keputusan pendidikan yang telah diambil berdasarkan hasil tes. Pengembang tes dapat membuat *data base* mengenai butir-butir soal yang bias tersebut sebagai bahan pijakan untuk mengembangkan tes di masa berikutnya agar dapat terbebas dari *DIF*.

Jika pada suatu riset tertentu, yang mungkin dilakukan oleh lembaga di luar pengembang tes, ditemukan adanya *DIF* pada suatu tes, maka menjadi kewajiban pengembang tes untuk melakukan tindak lanjut terhadap adanya *DIF* pada tes yang dikembangkannya. Hal ini sesuai dengan prosedur dalam testing yang mengatakan bahwa: "*when credible research reports that differential item functioning exists across age, gender, racial/ethnic, cultural, disability, and/or linguistic groups in the population of test takers in the domain measured by test, test developers should conduct appropriate studies when feasible*" (*AERA, APA, & NCME*, 1999: 81). Ini berarti, pendeteksian *DIF* dapat dilakukan oleh lembaga atau pihak-pihak di luar pengembang tes dan menjadi kewajiban pengembang tes untuk memperhatikan dan menindaklanjuti temuannya.

METODE PENDETEKSIAN DIF

Ada beberapa metode pendeteksian *DIF* yang telah dikembangkan oleh para pakar. Menurut Scheuneman dan Bleistein (1999: 220), metode pendeteksian *DIF* terbagi menjadi dua kelompok besar,

yaitu metode yang berdasarkan pendekatan teori tes klasik dan metode yang berdasarkan pendekatan teori respons butir. Perbedaan utama di antara keduanya adalah sebagai berikut. Metode pendeteksian *DIF* berdasarkan pendekatan teori tes klasik menggunakan skor total terobservasi (*total observed score*) sebagai representasi kemampuan peserta tes. Di sisi lain, pada metode pendeteksian *DIF* berdasarkan pendekatan teori respons butir, kemampuan peserta tes ditunjukkan oleh variabel laten (*latent variable*) dan karakteristik butir soal dinyatakan dalam parameter butir soal yang diestimasi berdasarkan teori respons butir. Metode-metode berdasarkan pendekatan variabel laten sering disebut metode parametrik, sebab menggunakan estimasi parameter, sedangkan metode-metode berdasarkan skor terobservasi sering disebut metode non-parametrik, sebab metode-metode tersebut tidak mengunakan model matematik untuk melakukan estimasi parameter (Penfield & Lam, 2000: 10).

Walaupun metode pendeteksian *DIF* berdasarkan teori respons butir lebih disukai dibandingkan dengan metode pendeteksian *DIF* berdasarkan teori tes klasik, karena landasan teorinya (Scheuneman dan Bleistein, 1999: 229), metode yang umumnya digunakan bukanlah metode yang berdasarkan teori respons butir. Metode yang justru banyak digunakan adalah metode berdasarkan teori tes klasik, misalnya metode Mantel-Haenszel (Embretson & Reise, 2000: 251), serta metode SIBTEST (Gierl, Khaliq, & Boughton, 1999: 10).

Pada pelaksanaan pendeteksian *DIF*, kelompok yang diselidiki apakah ada butir yang bias padanya disebut kelompok fokus (*focal group*) dan kelompok pembandingnya disebut kelompok acuan (*reference group*). Di Amerika Serikat, misalnya, biasanya yang ditentukan sebagai kelompok acuan adalah kelompok kulit putih, sedangkan yang ditentukan sebagai kelompok fokus adalah kelompok kulit hitam. Dalam perspektif gender, kelompok perempuan dapat ditentukan sebagai kelompok fokus dan kelompok acuannya adalah kelompok laki-laki, atau sebaliknya.

Berikut ini disajikan secara singkat salah satu metode pendeteksian DIF yaitu metode Mantel-Haenszel.

METODE MANTEL-HAENSZEL

Pada tahun 1959, Mantel dan Haenszel menampilkan prosedur untuk suatu studi pemadanan kelompok, yang oleh Holland dan

Thayer (1988: 129) dipakai untuk mendeteksi *DIF*, yang kemudian terkenal dengan metode Mantel-Haenszel. Metode ini merupakan metode yang *powerful* dan digunakan di *Educational Testing Service* (ETS) di Amerika Serikat (Dorans & Holland, 1993: 38).

Pada penggunaan metode Mantel-Haenszel, peserta tes pada masing-masing kelompok (kelompok fokus dan kelompok acuan) digolongkan menjadi M buah kategori berdasarkan pada level kemampuan peserta tes. Kemampuan peserta tes ini disebut variabel pemadanan (*matching variable*), yaitu variabel yang dipakai sebagai dasar untuk pemadanan (*matching*) (Holland & Thayer, 1993: 39). Pada metode Mantel-Haenszel, kemampuan peserta tes diwakili oleh skor total peserta tes. Menurut Holland dan Thayer, pengeluaran butir soal yang diselidiki *DIF*nya dari variabel pemadanan menyebabkan prosedur pendeteksian Mantel-Haenszel "*wili not behave correctly when there is no DIF*" (Dorans & Holland, 1993:60). Oleh karena itu, pada metode Mantel-Haenszel, variabel pemadanan harus memuat butir soal yang diselidiki *DIF*nya.

Tabel 11.1. Tabel Kontingensi 2 × 2 untuk Butir Soal Tertentu pada Level Kemampuan ke-m

	Banyaknya Peserta Tes yang Menjawab Benar	Banyaknya Peserta Tes yang Menjawab Salah	Banyaknya Peserta Tes Secara Keseluruhan
Kelompok fokus (f)	R_{fm}	W_{fm}	N_{fm}
Kelompok acuan (r)	R_{rm}	W_{rm}	N_{rm}
Kelompok total (t)	R_{tm}	W_{tm}	N_{tm}

Data yang digunakan dalam metode Mantel-Haenszel adalah data pada tabel kontingensi 2×2 sebanyak M buah atau data pada sebuah tabel kontingensi besar berukuran 2×2×M, dengan M adalah banyaknya penggolongan atas dasar level kemampuan peserta tes. Setiap tabel kontingensi 2×2 berbentuk seperti pada Tabel 11.1.

Hipotesis nol *DIF* pada metode Mantel-Haenszel yang diuji secara statistik adalah:

$$H_0 : \frac{\frac{R_{rm}}{W_{rm}}}{\frac{R_{fm}}{W_{fm}}} = 1, \text{ untuk } m = 1, 2, 3, \ldots, M \qquad 11.1$$

atau

$$H_0 : \frac{R_{rm}}{W_{rm}} = \frac{R_{fm}}{W_{fm}}, \text{ untuk } m = 1, 2, 3, \ldots, M \qquad 11.2$$

Hipotesis nol tersebut menyatakan bahwa *odds* untuk mendapatkan jawaban benar pada kelompok fokus dan kelompok acuan adalah sama pada setiap level variabel pemadanan.

Mantel dan Haenszel (Holland & Thayer, 1988: 133) mengembangkan tes khi-kuadrat dari hipotesis nol *DIF* melawan hipotesis alternatif, yang disebut *constant odds ratio hypothesis*, yang dirumuskan sebagai berikut.

$$H_a : \frac{R_{rm}}{W_{rm}} = \alpha \frac{R_{fm}}{W_{fm}}, \text{ untuk } m = 1, 2, 3, \ldots, M \text{ dan } \alpha \neq 1 \qquad 11.3$$

Perhatikan bahwa jika $\alpha = 1$, hipotesis alternatif tersebut berubah menjadi hipotesis nol *DIF* pada Persamaan 11.1. Parameter α disebut *common odds ratio* pada M buah tabel kontingensi 2×2, sebab di bawah H_a, nilai α adalah *common odds* untuk setiap m, yaitu:

$$\alpha_m = \frac{\frac{R_{rm}}{W_{rm}}}{\frac{R_{fm}}{W_{fm}}} = \frac{R_{rm}W_{fm}}{R_{fm}W_{rm}} \qquad 11.4$$

Mantel dan Haenszel menyediakan estimasi untuk *common odds ratio* sebagai berikut (Holland & Thayer, 1988: 134; Dorans & Holland, 1993: 40):

$$\hat{\alpha}_{MH} = \frac{\sum_m \frac{R_{rm}W_{fm}}{N_{tm}}}{\sum_m \frac{R_{fm}W_{rm}}{N_{tm}}} \qquad 11.5$$

Estimasi tersebut adalah estimasi ukuran efek (*efect size*) *DIF* pada metrik yang rentangannya mulai dari 0 sampai dengan ∞, dengan nilai 1 mengindikasikan adanya *DIF* yang nol, artinya butir soal yang bersangkutan tidak memuat *DIF*. Jika $\hat{\alpha}_{MH} > 1$, maka butir yang diselidiki lebih menguntungkan kelompok acuan. Jika $\hat{\alpha}_{MH} < 1$, maka butir yang diselidiki lebih menguntungkan kelompok fokus.

Uji Signifikansi

Mantel dan Haenszel mengembangkan statistik tes khi-kuadrat untuk menguji signifikansi hipotesis nol $H_0: \alpha_m = 1$, untuk setiap m. Statistik tes khi-kuadrat tersebut dirumuskan sebagai berikut (Holland & Thayer, 1988: 134; Dorans & Holland, 1993: 40):

$$MH_{\chi^2} = \frac{\left[\left|\sum_m R_{rm} - \sum_m E(R_{rm})\right| - 0{,}5\right]^2}{\sum_m Var(R_{rm})} \qquad 11.6$$

dengan

$$E(R_{rm}) = E(R_{rm} \mid \alpha = 1) = \frac{N_{rm}R_{tm}}{N_{tm}}$$

$$Var(R_{rm}) = Var(R_{rm} \mid \alpha = 1) = \frac{N_{rm}R_{tm}N_{fm}W_{tm}}{N_{tm}^2(N_{tm}-1)}$$

Statistik uji MH_{χ^2} pada Persamaan 11.6 berdistribusi khi-kuadrat dengan derajat kebebasan 1, jika H_0 benar. Kriteria pengambilan keputusannya adalah sebagai berikut. Jika $MH_{\chi^2_{obs}} > \chi^2_{\alpha:1}$, maka butir soal yang bersangkutan secara signifikan terdeteksi *DIF*.

Contoh 11.1

Misalnya terdapat 10 butir soal dengan 40 peserta tes (20 peserta kelompok acuan dan 20 peserta kelompok fokus) seperti yang tampak pada Tabel 11.2. Sebagai kelompok acuan, misalnya, kelompok siswa laki-laki dan sebagai kelompok fokus, misalnya, kelompok siswa perempuan.

Tabel 11.2. Sebaran Skor untuk 20 Siswa Kelompok Acuan dan 20 Siswa Kelompok Fokus pada 10 Butir Soal

No urut Siswa	Kelompok Acuan						Kelompok Fokus					
	Btr 1	Btr 2	Btr 3	...	Btr 10	Skor Total	Btr 1	Btr 2	Btr 3	...	Btr 10	**Skor Total**
1	1	1	1		0	9	1	1	1		0	9
2	1	0	1		1	9	0	1	1		1	9
3	1	1	1		1	9	1	1	1		1	9
4	1	1	1		1	9	1	1	1		1	9
5	1	1	1		1	9	1	1	1		1	9
6	1	1	1		1	9	1	1	1		1	9
7	1	0	1		1	8	0	1	1		1	8
8	1	1	0		1	8	1	1	0		1	8
9	1	1	1		0	7	1	1	1		1	7
10	1	1	0		1	7	1	1	0		1	7
11	1	1	1		1	6	1	1	1		0	6
12	1	0	0		0	6	0	1	0		1	6
13	1	0	1		0	5	0	1	1		1	5
14	1	0	0		1	5	0	1	0		0	5
15	1	0	1		0	4	0	1	1		0	4
16	1	0	0		1	4	0	1	0		0	4
17	0	1	1		0	3	1	0	1		0	3
18	1	0	0		0	3	0	1	0		0	3
19	1	1	0		0	2	1	0	0		1	2
20	0	0	0		1	2	0	1	0		0	2

Penghitungan indeks DIF dan Uji Signifikansi untuk Butir Soal Contoh 11.1 adalah sebagai berikut.

Skor Total	Klp Acuan		Klp Fokus					
	Rr	Wr	Rf	Wf	RrWf/Nt	RfWr/Nt	E(Rr)	Var(Rr)
9	6	0	5	1	0,5	0	5,5	0,25
8	2	0	1	1	0,5	0	1,5	0,25
7	2	0	2	0	0	0	2	0
6	2	0	1	1	0,5	0	1,5	0,25
5	2	0	0	2	1	0	1	0,333
4	2	0	0	2	1	0	1	0,333
3	1	1	1	1	0,25	0,25	1	0,333
2	1	1	1	1	0,25	0,25	1	0,333
Jumlah	18	2	11	9	4	0,5	14,5	2,083

$$\hat{\alpha}_{MH} = \frac{\sum_m \frac{R_{rm} W_{fm}}{N_{tm}}}{\sum_m \frac{R_{fm} W_{rm}}{N_{tm}}} = \frac{4}{0.5} = 8$$

$$MH_{\chi^2} = \frac{\left[\left|\sum_m R_{rm} - \sum_m E(R_{rm})\right| - 0,5\right]^2}{\sum_m Var(R_{rm})} = \frac{[|18-14,5|-0,5]^2}{2,083} = 4,320$$

Diperoleh $\hat{\alpha}_{MH} = 8$ dengan $MH_{\chi^2} = 4,320$. Jika diambil tingkat signifikansi $\alpha = 5\%$, yang berarti $\chi^2_{0,05;1} = 3,841$, maka butir soal nomor 1 signifikan terkena *DIF*. Karena $\hat{\alpha}_{MH} > 1$, maka butir soal nomor 1 terkena *DIF* yang menguntungkan kelompok acuan.

BAHAN DISKUSI

1. Ketika memvalidasi kisi-kisi pada validasi pakar, pada aspek bahasa, sering ditulis indikator "Tidak menggunakan bahasa daerah tertentu (setempat)". Apakah memasukkan indikator itu termasuk usaha pengurangan bias butir? Jelaskan alasan Anda.

2. Perhatikan butir soal berikut.

   Ayah menyembelih seekor sapi untuk dijual dagingnya. Berat dagingnya adalah 500 kg. Jika harga daging adalah Rp20.000,00 per kg, berapa uang yang diperoleh ayah jika semua daging dijual?

   Apakah butir soal seperti itu membuat siswa daerah tertentu merasa terusik jiwanya, mengingat pada daerah itu sapi termasuk hewan keramat? Apakah butir soal seperti itu termasuk butir soal yang bias?

3. Perhatikan butir soal berikut.

   Terdapat 3 buah becak dan 1 buah gerobag sampah. Berapa jumlah rodanya?

   Apakah butir soal itu merugikan siswa daerah tertentu? Mengapa?

4. Tulislah sebuah butir soal yang memungkinkan siswa daerah tertentu tidak dapat mengerjakan karena ada istilah yang tidak dimengerti. Apakah butir soal seperti itu bisa diujikan untuk Ujian Nasional? Mengapa?

5. Perhatikan data pada Tabel 11.2. Lakukan analisis DIF untuk butir soal nomor 2, 3, dan 10.

***

DAFTAR PUSTAKA

Allen, M. J. dan Yen, W. M. 1979. *Introduction to measurement theory*. California: Brook/Cole Publishing Company.

Anderson, L. W. 1981. *Assessing affective characteristics in the schools*. Boston: Allyn and Bacon. Inc.

Anderson, O.W. & Krathwohl, D.R. 2001. *A taxonomy for learning, teaching, and assessing: A Revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. Dalam P. W. Holland & H. Wainer (Eds), *Differential Item Functioning*. (pp. 3-23). Hillsdale: Lawrence Erlbaum Associates Publisher.

Asmawi Zainul dan Noehl Nasution. 1995. *Penilaian hasil belajar*. Jakarta: Direktorat Jenderal Pendidikan Tinggi, Departemen Pendidikan dan Kebudayaan.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). 1999. *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Barret, S. 2001. Differential item functioning: A case study from first year economics. *International education journal*, *2*, 123-132. Diambil pada tanggal 10 Februari 2003, dari http://www.flinders.edu.au/education/iej.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. *Assessment for learning: Putting it into practice*. Berkshire: Open-University Press.

Callison, D. 1988. Authentic assessment. *School library media activities monthly*. 14 (5). 1 – 4.

Camilli, S. & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks: Sage Publications.

Chance, B. L. 1997. Experiences with authentic assessment techniques in an introductory statistics course. *Journal of statistics education*. 5(3). Diambil dari https://www.amstat.org/publications/jse/v5n3/chance.html pada 1 Mei 2010.

Clarke, S. 2005. *Formative assessment in the secondary classroom*. London: Hodder Murray.

Crocker, L dan Algina, J. 1986. *Introduction to classical and modern test theory*. New York: CBS College Publishing

Dali S. Naga (1992). *Pengantar teori skor pada pengukuran pendidikan*. Jakarta: Gunadarma.

DiRanna, et. al. 2008. *Assessment-centered teaching: A reflective practice*. Thousand Oaks: Corwin Press

de Lange, J. 1999. *Framework for classroom assessment in mathematics*. TK: Freudental Institute & National Center for Improving Student Learning and Achievement in Mathematics and Science. Diambil dari http://scholar.google.co.id/scholar pada 2 Mei 2010.

Dorans, N. J. & Holland, P. W. 1993. DIF detection and description: Mantel-Haenszel and standardization. Dalam P. W. Holland & H. Wainer (Eds), *Differential Item Functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates Publisher.

Djemari Mardapi, dkk. 2002. *Pola induk sistem pengujian hasil kegiatan pembelajaran berbasis kemampuan dasar sekolah menengah umum (SMU)*. Jakarta: Depdiknas.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Marwah, NJ: Lawrence Erlbaum Associates Publisher.

French, A. W. & Miller, T. R. 1996. Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*, 315-332.

Gable, R. K. 1986. *Instrument development in the affective domain*. Boston: Kluwer-Nijhoff Publishing.

Garfield, J. B. 1994. Beyond testing and gradings: using assessment to improve student learning. *Journal of statistics education*. 2 (1). Diambil dari https://www.amstat.org/publications/jse/v2n1/garfield.html pada 1 Mei 2010.

Gierl, M. J., Khaliq, S. N. & Boughton, K. 1999. Gender differential item functioning in mathematics and science: Prevalence and policy implications. *Paper*. Presented at the annual meeting of the Canadian society for the study of education. Diambil pada tanggal 20 Januari 2003, dari http:// www.ncrel.org/sdrs.

Gronlund, N. E. 1985. *Measurement and evaluation in teaching*. New York: Macmillan Publishing Company.

Guilford, J. P. 1954. *Psychometric methods*. New York: McGraw-Hill Book Company, Inc.

Hall, C.S., Lidsey, G. & Campbell, J.B. 1997. *Theories of personality* (fourth edition) . New York: John-Willey.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park: Sage Publications.

Holland, P. W. & Thayer, D. T. 1988. Differential item performance and the Mantel-Haenszel procedure. Dalam H. Wainer & H. I. Braun (Eds), *Test validity* (pp 129-145). Hillsdale: Lawrence Erlbaum Associates Publisher.

Hulin, C. L., Drasgow, F. & Parson, C. K. 1983. *Item response theory: Application to psychological measurement*. Homewood: Dow Jones-Irwin.

Jahja Umar, Herwindo Haribowo, Bahrul Hayat, & Abdul Manan Akhmad. 1998. *Bahan penataran pengujian pendidikan*. Jakarta: Pusat Penelitian dan Pengembangan Pendidikan, Balitbangdikbud, Departemen Pendidikan dan Kebudayaan.

Johnson. D.W. & Johnson, R. T. 2002. *Meaningful assessment*. Boston, MA: Allyn & Bacon.

Kane. M. T. 2001. Current Concerns in Validity Theory. *Journal of educational measurement*. 38 (4).

Mansyur. 2009. *Pengembangan model assessment for learning pada pembelajaran Matematika di SMP*. Disertasi, tidak diterbitkan. Yogyakarta: Universitas Negeri Yogyakarta.

Moon, T.R., Brighton, C.M., Callahan, C.M., & Robinson, A. 2005. Development of authentic assessments for the midle school classroom. *The Journal of secondary gifted Education*. 16 (2/3). 119 – 135.

Mueller, J. 2.005. The Authentic Assessment Toolbox: Enhancing Student Learning through Online Faculty Development. *Journal of Online Learning and Teaching*. 1 (1).

Messick, S. 1989. Validity. *Educational measurement*: Third Edition. Edited by R.L. Linn. New York: Macmillan Publishing Company.

Miller, M.D., Linn, R.L.& Gronlud, N.E. 2009. *Measurement and Assessment in teaching: Tenth edition*. New Jersey: Pearson.

Nuning Hidayah Sunani. 2010. *Sistem penilaian berbasis kelas dalam pembelajaran Bahasa Indonesia: Studi kebijakan di SMP negeri Kabupaten Karanganyar*. Disertasi, tidak diterbitkan. Surakarta: Program Pascasarjana UNS.

Nunnally, J. C. 1978. *Psychometric theory*. New Delhi: Tata McGraw-Hill Publishing Company Limited.

Penfield, R. D. & Lam, T. C. M. 2000. Assessing differential item functioning in performance assessment: Review and recommendations. *Educational measurement: Issues and practice, 19*, 5-15.

Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 66 Tahun 2013 tentang Standar Penilaian Pendidikan

Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 104 Tahun 2014 tentang Penilaian Hasil Belajar oleh Pendidik pada Pendidikan Dasar dan Menengah.